Escola de Camins
Escola Tècnica Superior d'Enginyeria de Camins, Canals i Ports
UPC BARCELONATECH

AN ANALYSIS OF BICING MOBILITY PATTERNS USING BIG DATA

TREBALL FINAL DE MÀSTER

Treball realitzat per:
Oriol Manchón Contreras

Dirigit per:
Marino Arroyo Balaguer
Behrooz Hashemian

Màster en:
Enginyeria de Camins, Canals i Ports

Barcelona, 23 de Juny del 2016

Departament de Matemàtica Aplicada III

Escola de Camins
Escola Tècnica Superior d'Enginyeria de Camins, Canals i Ports
UPC BARCELONATECH

Màster en Enginyeria de Camins, Canals i Ports

Treball Final de Màster

An analysis of Bicing Mobility Patterns using Big Data

Autor:
Oriol Manchón Contreras
Supervisors:
Marino Arroyo Balaguer
Behrooz Hashemian

June 22, 2016

This page is intentionally left blank.

An analysis of Bicing mobility patterns using Big Data

Oriol Manchón Contreras
Màster en Enginyeria de Camins, Canals i Ports, 2016
Professor: Marino Arroyo Balaguer,
Department of Civil and Environmental Engineering, UPC
External supervisor: Behrooz Hashemian,
Department of Urban Studies and Planning, MIT.

Nowadays, technology advances really fast and so does the generation of data. Almost all electronic devices are constantly generating and sharing a huge amount of data through the World Wide Web. Moreover, recent policies of open governments and data, are helping to make available this information for everybody that wants to take it and use it. The aim of using Big Data is to discover knowledge that is hidden behind thousands of rows of information. However, to find out the value of the data, it is necessary to use non-traditional methods able to deal with such amount of information.

Furthermore, big cities have traffic problems and complex mobility patterns which need to be studied in depth to improve life conditions of citizens, reduce pollution and to create eco-friendly cities. This work is focused on the city of Barcelona and its bike-sharing system Bicing. The aim is to understand the mobility patterns of Bicing subscribers using Big Data.

Treating Big Data requires of more resources than conventional problems. So that, setting a methodology to acquire, pre-process and treat the data has been necessary before proceeding with the analysis.

In order to gain visibility out of the data, two different approaches have been followed. First of all, an exploratory analysis of the behaviour of the users of Bicing. On the other hand, a Principal Component Analysis has also been carried out to understand the data but also to reduce the dimensionality, hence the volume of the data necessary to provide acceptable results.

To sum up, the present work is a particular example of the possibilities that Big Data offers in terms of gaining knowledge out of massive amounts of data. Moreover, it studies the patterns of Bicing subscribers during different periods of the day, week and year based on real data.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Prof. Marino Arroyo of Department of Civil and Environmental Engineering at Universitat Politècnica de Catalunya. The door to Prof. Arroyo office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my own work, but steered me in the right direction whenever he thought I needed it.

I would also like to thank Dr. Behrooz Hashemian, Department of Urban Studies and Planning at MIT, who guided me since the first moment through all this work. Each time I was in need of help and support, he was there to help me see things clear. Without his passionate participation and input, this thesis could not have been successfully conducted.

I would also like to acknowledge David Ortin of the IT Department at Universitat Politècnica de Catalunya, who gave me access to the computer laboratory and research facilities, and helped me with software and servers problems.

I would like to thank as well Gabriel Martins, Boris Bellalta and Simon Oechsner, students at Universitat Pompeu Fabra who kindly shared the Bicing data with us. Without them the dataset used for this thesis would have been much smaller and getting results for an entire year would not have been possible.

Finally, I must express my very profound gratitude to my parents, friends and to my girlfriend for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you.

Contents

List of Figures

List of Tables

Listings

1 INTRODUCTION

1.1 Context

As technology advances, data gathering is becoming much easier and cheaper, and when this happens, all the data stored also becomes more accessible for everybody. However, this data requires an analysis to find out what is hidden behind it, in other words, to uncover the hidden value of this data and use it. If the data is analysed properly, the applications are limitless and either the business world or citizens can take some advantage out of it. The present work is a particular case of the aforementioned but before starting, it is necessary to provide some background.

1.1.1 Big Data

First of all, it is necessary to understand what is Big Data. Nowadays, Big Data is being generated by everything around us at all times. Every digital process and social media exchange produce it. Systems, sensors and mobile devices transmit it. A formal definition for big data was given by Gartner analyst Doug Laney who said in 2011 that, "*Big Data*" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making [2].

Figure 1: Overview of Big Data – *DNP Analytics*

The original definition has been completed with other dimensions, such as Variability, Veracity, Visualization and Value, in order to provide a more complete insight on what Big Data is. Each of these seven dimensions stands for:

- **Volume**: It refers to the vast amounts of data generated every second. This makes most data sets too large to store and analyse using traditional database technology. New big data tools

use distributed systems so that the users can store and analyse data across databases that are dotted around anywhere in the world.

- **Velocity**: This term refers to the speed at which new data is generated and the speed at which data moves around. Technology allows now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases. Hence, the Velocity is the speed at which the data is created, stored, analysed and visualized.

- **Variety**: It refers to the different types of data that can now be used. In the past, analysts were only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unstructured (text, images, video, voice and so on). However, with Big Data technology data of different types can be analysed and brought together to provide a better understanding of what is happening.

- **Variability**: The intrinsic meanings and interpretations of these conglomerations of raw data depend on its context.This is especially true with natural language processing. A single word may have multiple meanings. New meanings are created and old meanings discarded over time. Interpreting connotations is, for instance, essential to gauging and responding to social media buzz. The boundless variability of Big Data therefore presents a unique decoding challenge if one is to take advantage of its full value.

- **Veracity**: This term refers to the biases, noise and abnormality in data. Having a lot of data in different volumes coming in at high speed is worthless if that data is incorrect. Incorrect data can cause a lot of problems for organisations as well as for consumers. Therefore, organisations need to ensure that the data is correct as well as the analyses performed on the data are carried out properly.

- **Visualization**: This is the hard part of Big Data. Making all the vast amount of data comprehensible in a manner that is easy to understand and read. With the right analysis and visualizations, raw data can be put to use otherwise raw data remains essentially useless. In this particular case, visualization refers to complex graphs that include many variables of data while still remaining understandable and readable.

- **Value**: All the available data will create a lot of value for organisations, societies and consumers. However, the value is not in the data itself but in the analysis performed on that data and how the data is turned into information and eventually turned into knowledge. The value is how organisations will use that data to make informed decisions.

It is important to highlight the **Value** of the data for companies and governments. For instance, some companies in many industries are taking advantage on its competitors by investing on a field with an opportunity of major revenue. The discovered knowledge can be applied to information management, query processing, decision making, process control, and many other applications [1].

1.1.2 Big Data in Civil Engineering

As it has been previously explained, the information is generated every second and it can be gathered and stored for analysis. There are many applications for this data and amongst them there are Civil Engineering challenges. In this section, some of these applications are to be addressed in order to highlight their benefits for society and improvement of citizens daily life. Before starting with the applications, it is worth to mention that the analysis of Big Data is also known as data mining [3].

The different applications of Big Data in Civil Engineering are based on data collected from sensors, which are located at certain strategical points. These sensors are the responsible of collecting the data that will be studied and understood. Some examples of the aforementioned are:

- **Smart Grids**: Regulation and control of electricity and water grids. Using the different sensors readings, the electricity/water consumption, demand and quality can be predicted and, taking into account that, the production can be adjusted to better fit the expected demand and to control the system itself.
- **Monitoring of Infrastructure Health**: Data sensing and analysis can help in monitoring the conditions of an infrastructure both above and below the ground. Data mining can be used to predict the health of the infrastructure as well as possible correction mechanisms associated.
- **Natural Disaster Assessment and Emergency Response Planning**: High-resolution images of areas affected by a natural disaster can be mined to detect debris fields, obstruction of ingress routes, building damage detection and also to establish a response planning to help the affected area [4].
- **Mobility and Traffic Planning**: This is one of the most interesting applications. Sensors deployed all around cities are increasing every day. The information collected from such sensors can be applied, for instance, to regulate traffic lights periods in order improve the circulation or to change travel signs depending on the demand on that particular street or highway or even to predict the real-time occupancy of parking lots in the streets. There is a Catalan company called *Worldsensing* that, by using sensors installed in the traffic lights and the Wi-Fi signal of car drivers smartphones, is able to predict real-time mobility patterns inside Barcelona and suggest alternative routes to customers in order to avoid congestion and traffic jams.
- **Supply Chain**: If the data of a company, which has some Supply Chain processes, is mined and understood, it is possible to improve its efficiency, thus reducing costs. One example is Tesco's, that expects to reduce 100 million pounds of the expenses related to the Supply Chain process.

There are many applications of Big Data in Civil Engineering but the interesting thing is that are many more which are not being studied yet. Things are changing pretty fast, data is produced even faster and we, as Civil Engineers, need to start opening our minds to these new technologies rather than focusing only on the conventional applications.

Nowadays, construction of huge infrastructures is not as common as it was few years ago. The trend now is to do the maintenance of these structures and Big Data offers new possibilities of doing that. High-competitiveness between companies requires to use smartly the available resources and the use of Big Data might have a great impact on cost reduction. For instance, placing sensors to predict the future behaviour of structures might reduce these maintenance costs, hence one company

might deal with more projects at the same time with the same budget or can win a project by offering the best deal.

This project is focused on one particular application of Big Data in Barcelona.

1.1.3 Application in Barcelona

One of the major problems Barcelona is facing is the traffic congestion, which makes life uncomfortable for people, and the resultant pollution has a great impact on the environment. With the increasing prices of public transports and the congestion problems during peak hours, an alternative to traditional transport methods is needed.

Bicing, which is a bike-sharing system, was born to give a solution to that problem. The service, which has 96.715 subscribers, allows users to use public bikes to move around and they only have to pay a subscription fee each year (which is around 47€/year). The system consists on 428 bike stations (see Figure 2) where the users (previously subscribed) can borrow bikes for a 30 minute ride with no cost. Extra costs only apply when the rides are longer than 30 minutes. Nevertheless, in case the user has almost consumed the entire 30 minute period and the destination station is full, he gets 10 minutes extra to find another station with free slots.

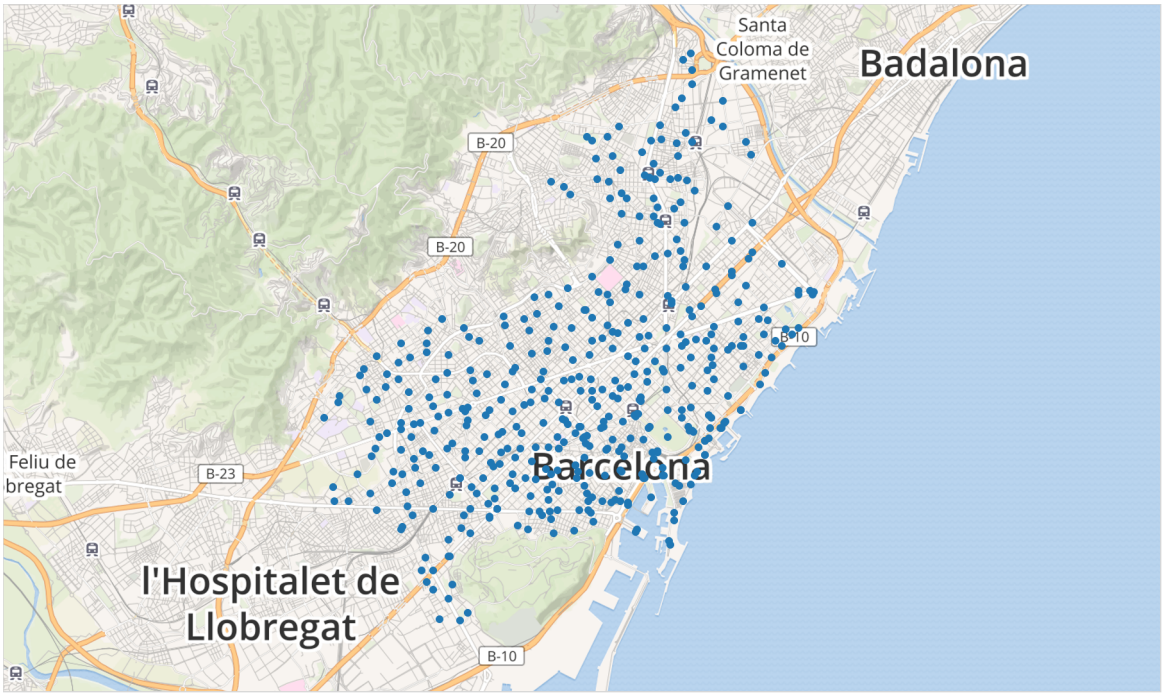

Figure 2: Bicing stations distributed all over the city of Barcelona

Regarding the maintenance of the system, it is carried out by the company itself. This maintenance does not only include taking care of the bikes but to transport the bikes around the city from full stations to empty ones using small but adapted transportation vehicles. Our first intuition is that, these vehicles are sent to the stations that are detected as full and the vehicle takes them to an empty station. This is not an instantaneous operation as the vehicles need to go from one station to another. Meanwhile, another stations of the same area can get full which means that the necessary time to fill (or empty) the stations can be quite long, which supposes a problem for the users.

Figure 3: Bicing station in Barcelona – *bicing.cat*

Using Bicing relieves the users of the responsibility of owning a bike, and the problems that imply. In fact, when using bike-sharing systems, users do not have to worry about spending extra money on maintenance or being worried about getting their bike stolen. Moreover, it is a eco-friendly transport method and promotes a healthier lifestyle which is a modern trend.

1.1.4 Other applications of Big Data

The aim of Big Data is also to improve decisions and competitiveness for companies and for the public administrations, which will create a significant growth of the world economy. Big Data helps to better listen to costumers, better understand their ways of using services and hence the offer. These applications also have benefits in many fields: from scientific researches to national security, and from global economy to society administration [5]. Following, there are the most widespread applications of Big Data problems:

- **Business & Commerce**: In the era of information, it has been shown that every 1.2 years, the volume of business data worldwide doubles, across almost companies. The insights hidden in this deluge of data can be discovered by using Bid Data, which help for example to optimize business processes, detect a pattern, better understand costumers, anticipate their behaviours, needs and intentions.

- **Science & Research**: The scientific domains are based on data-intensive scientific discovery. For instance, the Swiss nuclear physics lab with its Large Hadron Collider (the world's largest and most powerful particle accelerator) uses the power of thousands of computers, distributed across 150 data facilities worldwide, to process all the data it produces.

- **Health**: Public health and medicine are fields in which employing effectively the healthcare data deluge would give the right outcomes to the patient and reduce care cost. The computing power of Big Data allow us to mine entire DNA strings in minutes and will provide us

the possibility to discover, monitor, improve health aspects of every one and predict disease patterns.

- **Smart Cities**: Modern cities and countries are currently being transformed by the new possibilities Big Data brings. Sustainable economic development and high quality of life, with rise management of natural resources are the major objectives of cities. For example, in many cities, the transport infrastructure and utility processes are all joined up in order to turn into smart cities. This will help them to minimize jams and optimize traffic flows.

There are unlimited applications of Big Data to solve problems that few years back were not even taken into account. Besides the aforementioned applications, there is one particular case where Big Data analysis was applied. This curious case is the TV-Show "House of Cards" produced by Netflix.

First, much like Google, Netflix has a lot of data. Currently, it has many millions of customers worldwide and spreads a very wide net to collect data on them. In particular, the site also logs something it calls "user actions".

These include the times of day people watch horror films and when they watch humour. It also logs when users start and stop viewing, what they rewind to watch again, whether they watch on a TV or iPad and so on. It even looks at pirate film sites to determine what is trending.

With that data, analysts were able to see what kind of TV-Show the users might like, taking into account the actor, the director, the genres and seeing the correlation between them. They found that people like films starring Kevin Spacey or films directed by David Fincher. Hence, they thought users would like a TV-Show that combined both of them. By thinking that way, they finally came up with "House of Cards" as it is, which has been a tremendous success with millions of viewers worldwide.

1.2 Objectives

The objectives of this project are divided in two main parts. First of all, it is necessary to understand what Big Data is and the implications of working with such type of data. This goes from collecting the data to preparing it for the analysis. The other part, will be an analysis of the Bicing data. This analysis aims to find some useful information out of the data.

Both parts will be as follows:

1.2.1 Establish Methodology

Considering that this project requires the acquisition, treatment, preparation and analysis of Big Data, establishing a methodology is one of the most important parts of the present study. This process follows a sequential order as depicted in Figure 4.

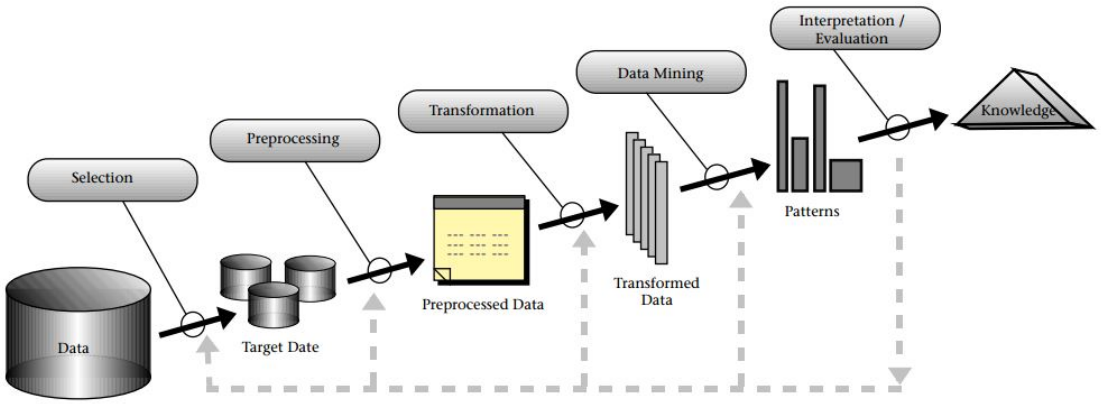

Figure 4: Different steps carried out on Big Data to find out the value of the data [1]

First of all, it is necessary to have a large dataset from which the target information (i.e records from Mondays to Fridays) is extracted. Afterwards, this data needs to be preprocessed in order to supply missing values and delete noisy data that may interfere in the analysis. Once the data is cleaned, it is necessary to transform it. This is, to select which are the features and dimensions that will be interesting to analysis. Next step is to apply the Data Mining algorithms (i.e associations, classification, clustering, etc.) on the data. To close the loop, the results need to be interpreted and evaluated to find the value of the data.

Taking into account the aforementioned, it will be necessary to set a procedure to acquire, treat, prepare and use the data before carrying out the analysis. In order to avoid problems of storage and have an easy access to the data, it will be required to use a server and a database management system (DBMS), which is a computer software application that interacts with the user, other applications, and the database itself to capture and retrieve data . Moreover, a potent software will be crucial to carry out that analysis and to represent the results.

The methodology part is explained with much more detail on Section 2.

1.2.2 Find usage patterns

After completing all the data acquisition and preparation of the data, an analysis will be performed. From this analysis, we expect to find some useful information hidden on the data, such as users behaviour or trends which are repeated daily, weekly, monthly or even between seasons. The analysis also intends to reduce dimensionality of the data and prove that most of the data can be explained with less dimensions. This might be useful to take decisions about future city planning, to improve the system and the user-experience or even to re-design the distribution of the stations over the whole city. This part is explained with much more detail on Section 3.

2 METHODOLOGY

2.1 Data acquisition

Due to open-government policies, the data produced is becoming free for everybody. Nevertheless, it might not be that easy to acquire an interesting dataset. Taking possession of a dataset is easy when the data is shared in open data websites of governments. Some of them might be, for instance, *https://www.data.gov/* (US Government's Open Data) or *http://opendata.bcn.cat/* (Barcelona Governemnt's Open Data) where users can find sets of data of one or more years that are stored there.

However, not all the datasets have been previously stored and ready to use. Normally, the data of a particular system is being generated every minute or every 10-minute period. If that data is not saved, it gets lost once the data is produced again. Consequently, it is necessary to automatically save it when it is generated by the system, before losing it. To do so, there is a technique known as web-scrapping which reads and stores the data automatically.

In this case, the Bicing data is generated by the system every minute and it is open for public access via an API (Application Programming Interface). The acquisition of the data is done using the aforementioned technique, which is "web-scrapping". This is a process of automatically collecting information from the World Wide Web. This technique focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analysed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software.

Using this technique, the data is extracted from the Bicing website and stored. One of the drawbacks of this method is that the collection of the aforementioned data is done in real-time (i.e. if we need a week of data, we have to program a script to web-scrap during a week non-stop). Consequently, unless previous data is stored and made public, it will be necessary to web-scrap for the desired period of time to obtain the data.

Nevertheless, in order to find useful value to the Bicing data, it is necessary to collect a large amount of it. To have one or two-week data, or even a couple of months of it, would not be sufficient to establish some usage patterns, or to find out the behaviour of the users of the system. Hence, at least a full year of data is necessary.

Since this project has been carried out in about six months, there was not enough time to gather all that information. However, it has been possible to acquire one year and a half worth of data thanks to the collaboration of UPF students; Gabriel Martins, Boris Bellalta and Simon Oechsner. They have been using Bicing data to work in their project [6] and were able to share this dataset with us.

2.2 Bicing data description

The size of the data file is about 17 Gigabytes. The starting date of the data sample is 13th of October 2013 and the last time is 14th of March 2016. This is more or less 1 year and 5 months of records. The data stored for a single week has approximately 5 millions rows. Considering one year and five months of data, the total number of rows goes up to **475 millions**. Each row of this data has six columns and each of these measurements (columns) stand for:

1. *Id*: Unique identification number assigned to all the stations each time the data is updated
2. *Update_time*: Time when data is acquired and stored
3. *Station_id*: Unique identification number for each station. There are 428 different stations.
4. *Status*: Status of the station, either Open (OPN) or Closed (CLS)
5. *Slots*: Number of available free spaces (or slots) of each station at a given time
6. *Bikes*: Number of available bikes of each station at a given time

A small sample of the data available in th dataset is depicted in Figure 5. As it can be seen the sample consists on a few number of rows but it is worth to remind that the whole table has 475 millions of rows.

id	update_time	station_id	status	slots	bikes
1454281262	2016-02-01 00:00:17.0	357	OPN	13	14
1454281262	2016-02-01 00:00:17.0	137	OPN	0	26
1454281262	2016-02-01 00:00:17.0	95	OPN	32	0
1454281262	2016-02-01 00:00:17.0	100	OPN	19	0
1454281262	2016-02-01 00:00:17.0	303	OPN	27	0
1454281262	2016-02-01 00:00:17.0	393	OPN	5	28
1454281262	2016-02-01 00:00:17.0	189	OPN	13	14
1454281262	2016-02-01 00:00:17.0	232	OPN	0	27
1454281262	2016-02-01 00:00:17.0	372	OPN	3	24
1454281262	2016-02-01 00:00:17.0	270	OPN	7	16
1454281262	2016-02-01 00:00:17.0	159	OPN	31	0
1454281262	2016-02-01 00:00:17.0	397	OPN	25	2
1454281262	2016-02-01 00:00:17.0	223	OPN	34	2
1454281262	2016-02-01 00:00:17.0	264	OPN	6	26
1454281262	2016-02-01 00:00:17.0	367	OPN	26	3
1454281262	2016-02-01 00:00:17.0	195	OPN	16	8
1454281262	2016-02-01 00:00:17.0	99	OPN	9	11
1454281262	2016-02-01 00:00:17.0	241	OPN	11	14
1454281262	2016-02-01 00:00:17.0	348	OPN	18	11
1454281262	2016-02-01 00:00:17.0	37	OPN	2	22
1454281262	2016-02-01 00:00:17.0	24	OPN	0	19
1454281262	2016-02-01 00:00:17.0	333	OPN	27	0
1454281262	2016-02-01 00:00:17.0	148	OPN	2	21

Figure 5: Sample of the Bicing Data

2.3 Software

In order to store, prepare and process the data, different software and tools have been used. In the present section, they are to be described and the relationship between their interfaces is also going to be addressed.

2.3.1 Servers: Lordvader and Learnfromdata

In Big Data projects, one of the important parts is the data access and storage. Since the size of the data is far larger than traditional datasets, it is necessary to store it in a server from where the data can be manipulated using different computers (if necessary).

For this particular project, two different servers have been used for that purpose. The main server is called Learnfromdata and it is a Virtual Machine (VM) running over a real server with a VMware ESXi system. Being a VM means that we can assign different resources depending on the project requirements. In this case, Learnfromdata has 4 CPU's and 8Gb RAM assigned. Furthermore, this server has MySQL client installed where we have the data stored and we run the different queries to retrieve the target data from the dataset. In order to access the server Learnfromdata, there are two different possibilities.

First of all, we can use any of the "Laboratori de Càlcul Numèric" (LaCaN) workstations. These are based on a multi-user Linux system with a shared disk called Lordvader. Hence, Lordvader is a disk server which allows users to store data and execute programs such as Matlab or LaTeX through a workstation.

On the other hand, in order to connect to Learnfromdata from the exterior, which is using a personal computer at home, it is necessary to connect using a SHH Tunnel through Lordvader. This procedure is only carried out for security purposes and it can be done using, for instance, a PuTTY client.

2.3.2 Database Management System (DBMS)

A database is a collection of information organised so that it can easily be accessed, managed and updated. In computing, databases are sometimes classified according to their organizational approach. The most prevalent one is the relational database, a tabular database in which data is defined so that it can be reorganised and accessed in a number of different ways.

Typically, a database manager provides users the capabilities of controlling read/write access, specifying report generation, and analysing usage. To do that, there is a standard language called SQL which stands for Structured Query Language. With this language it is possible to make interactive queries to extract information or update a database. There are many different database providers but the most common ones are IBM's *DB2*, Microsoft's *SQL Server*, and database products from Oracle, Sybase, and Computer Associates.

In this case, the database management system used is MySQL, which is an open-source relational database management system. MySQL was installed in Learnfromdata Virtual Machine and the "bicing" database was created with a table called "raw" which consists on the data set previously acquired. Since the preprocessing, treatment and analysis is carried out in MATLAB, only one query to retrieve the data between two dates is necessary. This query has the format shown in Listing 1.

Listing 1: MySQL Query used to retrieve the desired data from "Bicing" database

```
mysql> SELECT * FROM raw.bicing
    -> WHERE update_time > '2015-02-22'
    -> AND   update_time < '2015-02-23'
    -> LIMIT 5;
```

id	update_time	station_id	status	slots	bikes
1424615131	2015-02-22 00:00:17	384	OPN	17	9
1424615131	2015-02-22 00:00:17	97	OPN	7	12
1424615131	2015-02-22 00:00:17	350	OPN	31	0
1424615131	2015-02-22 00:00:17	325	OPN	16	8
1424615131	2015-02-22 00:00:17	426	OPN	12	19

```
5 rows in set (0.00 sec)
```

This query selects all the columns from the table "raw" inside "bicing" database where the date is between the range specified in the "update_time" variable. Note that a limit of 5 rows has been specified just to show the format of the query and of the data retrieved but in a real query this limit should not be set.

2.3.3 MATLAB

Considering that this project involves Big Data treatment and analysis, conventional software does not suit the requirements. For instance, the spreadsheets (i.e Microsoft Excel) do not allow tables with more than 4 million rows. The Bicing dataset has more than 475 millions of rows and it would be impossible to work with a spreadsheet. Consequently, a more sophisticated program is required to carry out this thesis. Amongst many other available software, MATLAB will be used for the present project.

MATLAB, which stands for "**MAT**rix **LAB**oratory", is a tool for numerical computation and visualization. A proprietary programming language developed by *MathWorks*, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and many more features.

Like other mathematical software, MATLAB has different built-in functions and add-ons or Toolboxes. Built-in functions, such as "mean", "std", or many others have been programmed by MathWorks analysts, hence it is not necessary to develop and test all the functions, making the experience simpler for users. Moreover, Toolboxes are a collection of routines that are designed to do common and not so common things. These are more complex than the simple normal programming syntax and built-in functions that the base MATLAB has. While the built-in functions are available for all users that have MATLAB License, Toolboxes require extra Licenses that need to be bought. Thanks to the "Universitat Politècnica de Catalunya" agreements with *MathWorks*, it is possible to use the Academic Version of MATLAB. It also offers the possibility to use different Toolboxes.

As it has been explained before, the Bicing dataset is stored in a MySQL server installed inside "Learnfromdata" server. For data manipulation either MySQL or MATLAB are valid options. However, for analysis purposes it is necessary to use MATLAB as it has the aforementioned built-in functions and Toolboxes.

For this particular project it will be necessary to use Database Toolbox (see Figure 6). This will allow us to exchange data between MySQL and MATLAB. Moreover, it will be possible to test the queries before retrieving the desired data as this Toolbox includes an App that allows to see small samples of the data.

In order to exchange data between MySQL and MATLAB, the connection needs to be configured previously. Once this is successfully set up, MATLAB is ready to start retrieving data.

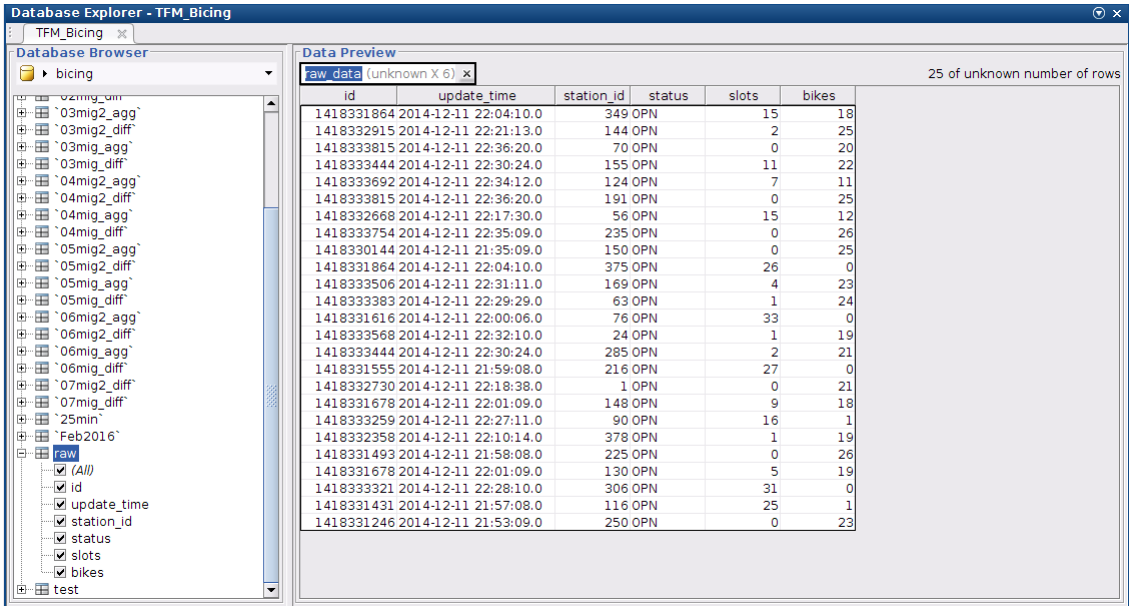

Figure 6: MATLAB Database Toolbox

Regarding the execution of the MySQL queries, these are to be done directly on MATLAB using the Database Toolbox. This toolbox allow users to introduce queries, check if they retrieve the appropriate data and store the query in MATLAB language. In order to retrieve the data from the database, the MATLAB script shown in Listing 2 is used. This script also transforms the "update_time" variable format to "datetime" in order to help grouping the data for analysis.

Listing 2: Script containing the query to retrieve the information from the database in MATLAB

```
1
2 function data = retrieve_bicing(StartTime, EndTime, StationID, TableSQL)
3
4 % Set default table
5 if ~exist('TableSQL', 'var')
6 TableSQL = 'raw';
7 end
8
```

```
% Set preferences
setdbprefs('DataReturnFormat', 'table');
setdbprefs('NullNumberRead', 'NaN');
setdbprefs('NullStringRead', 'null');


% Make connection to database using JDBC driver.
conn = database('bicing', 'user_name', 'user_password', 'Vendor', 'MYSQL
    ', 'Server', '127.0.0.1', 'PortNumber', 3306);

% Read data from database.
if exist('StationID','var')
data = fetch(conn, ['SELECT * '...
' FROM  bicing.', TableSQL, ...
' WHERE ',TableSQL,'.station_id = ', num2str(StationID)...
' AND   ',TableSQL,'.update_time > ''',StartTime,''''...
' AND   ',TableSQL,'.update_time < ''',EndTime,'''']);
else
data = fetch(conn, ['SELECT * '...
' FROM  bicing.', TableSQL, ...
' WHERE ',TableSQL,'.update_time > ''',StartTime,''''...
' AND   ',TableSQL,'.update_time < ''',EndTime,'''']);
end

% Close database connection.
close(conn);

% Change the date string to datetime format
data.update_time = datetime(data.update_time, 'InputFormat','yyyy-MM-dd
    HH:mm:ss.S');
```

2.4 Data preparation

In this section, which is divided in two different parts, and the procedures carried out to prepare the data for analysis are introduced and briefly discussed.

2.4.1 Data preprocessing

One of the problems of big datasets is that they are suitable to have missing values. Depending on the analysis method that is being performed, having missing data in the table could compromise the results or stop the analysis as the required operations cannot be done. In order to avoid this kind of problems when performing the analysis, the data needs to be preprocessed.

Data cleaning and preprocessing includes basic operations, such as removing noise or outliers, collecting the necessary information to model or account for noise and deciding on strategies for handling missing data fields.

In many cases, the missing data affects a singular row or just a couple of values in some rows. To decide what to do with these values, it is necessary to understand the data that we have and take a decision. For instance, if there is an entire row with missing values, that row would normally be deleted but, if it is just one value that misses, it might be replaced by the previous one. These kind of actions need to consider whether the deletion or modification of the data can affect the outcome or not.

In this particular project, the data is recorded every minute which means that, if some values are missing, those rows can either be modified or deleted as one minute is a small time period compared with the whole year that will be analysed. However, it seems more comprehensible to replace the data rather than erase it because the missing values are highly probable to be the same as the previous ones. So, for the methods that require matrices with non-missing values, a replacement has been done. Moreover, to reduce the computational cost, some of the analysis have been done making aggregations of the data, that is, grouping the data in time intervals, i.e. 10 minute-interval.

The considered period for this analysis has been the entire 2015, and the total data after cleaning it and removing days with many missing registers consists on 419 stations and 342 days (out of 365 days that a year has).

2.4.2 Data transformation

Analysing 475 million rows all together will consume an enormous amount of computational resources as well as time. In order to carry out an efficient analysis and achieve reliable results it will be necessary to previously transform the data.

Considering that the data is generated each minute, it seems reasonable to aggregate the data hourly and replace the exact number of bikes with the average and the standard deviation for each station. The interesting value in this case is the standard deviation obtained hourly for each station, which measures the activity at each hour. Hence, the larger is the value, the greater the activity of the station.

Following this procedure, one day of data that initially had 600.000 rows now has only 10.000 rows. Consequently, one entire year of data has 365.000 rows which, in comparison with the initial size of the data, which was more or less 250 million rows, is a reasonable value for analysis. It needs to be taken into account that now we have two different tables of 365.00 rows each but it is still a good value to work with in terms of computational cost.

However, the transformation of the data does not only aim to reduce the rows of the original table. At the beginning, the dataset had six columns and millions of rows. The interesting variables for analysis of this data are the "update_time", the "station_id" and the "number of bikes". Hence, the original raw data (see Figure 5) is to be transformed in a table with:

- **Rows**: These is equal to the total number of hours considered in the analysis period (i.e for five days there would be 120 rows. In this case, 342 days of analysis are 8208 rows)
- **Columns**: These is equal to the total number of stations in the system plus one. The first column consists on the daily hours for each day of the table (in this analysis, only 419 stations are considered, hence the table has 420 columns)

Consequently, the data to be analysed has the format depicted in Figure 7, which shows a sample of the activity (standard deviation) at the stations. However, in Figure 7 only few stations and hours can be appreciated but the total dimension of the table is 8208 rows and 420 columns.

	1 hGroup	2 x1	3 x3	4 x4	5 x5	6 x6	7 x7	8 x8	9 x9	10 x10	11 x11	12 x12	13 x13	14 x14
1	0	4.3390	2.3898	0.4576	11	17.8136	26.5254	22.6610	6.5763	6.9322	4.9831	16.6780	9.8644	8
2	1	4.0862	2.3621	0.0862	11	14.3621	26.2241	23.1379	6	6.3276	4	19.5345	9.7931	7.3103
3	2	1.0169	0.1864	0.3220	10.5085	12.8814	25.4746	23.6780	8	8.6949	3.2034	21.6949	4.7627	7
4	3	1	0.9138	0.7069	10	12	24	23.9138	7.0345	9.8448	2.4483	22.9483	3.8103	7
5	4	0.8966	0.5517	0.8103	10	10.0345	24	24	7	10.8103	2	24.9655	5	7
6	5	1.1552	2.0690	0.7586	10	11.0345	23.7759	23.3448	7	10.1207	2	29	5.5345	6.2931
7	6	3.9483	6.0172	0.9138	10	10.2759	25.5862	24.9483	7	10	2	29.5690	6.8448	5.1207
8	7	3.8966	4.4483	0.0172	9.3966	9.5862	26.7759	25.0345	7	10	2	29.6207	6.5517	4.8621
9	8	1.9492	4.0339	0.5593	9.0508	9.3220	27	25.9661	7	10	2.7288	28.1695	6.1695	4.9322
10	9	1.5000	5.3966	1.0690	21	21.0517	27	26	6.9483	10	3	26.2759	1.6724	5
11	10	2.6441	6.4576	1.0847	19.6949	24.3390	27	26.4237	7.1864	9.2542	2.5932	24.1017	0.2373	4.4407
12	11	2.4310	5.9828	3.1034	19.6552	24.8103	27	27	6.6897	4.9138	1.6034	17.3276	0	5.0345
13	12	2.0169	8.1864	2.1525	19.4407	24.8136	26.5254	26.9153	4.2881	1.8305	2.6441	6.3390	1.2881	4.4068
14	13	1.9655	7.3448	0.7414	18.1379	24.2586	24.6207	26.2931	2.5345	1.1724	4.9483	2.3276	11.7414	4.7586
15	14	1.6034	4.7586	1.6379	16.9828	24.6552	22.0345	20.0517	1.9310	5.6207	10.7759	1.6724	8.9138	1.6724
16	15	2.5000	0.8966	2	18.8793	22.6552	17.3966	20.5000	0.1207	4.5862	12	1.1034	4.6207	0.3448
17	16	4.6207	1.9828	2.1724	33.5690	31.7069	25.1724	24.8621	13.8448	4.0862	10.8276	0.7586	2.8621	8.5345
18	17	4.7759	2.3793	1.5517	31.3621	32.9310	23.0690	22.1379	17.0172	4.4138	11.8448	3.6552	4.9138	18.9138

Figure 7: Sample of the data after being cleaned and reshaped, ready for analysis

The data shown in Figure 7 has been grouped by hours. With this data, an analysis on the activity depending on the day-time can be carried out. Nevertheless, it is also interesting in terms of analysis to group that data by week-days and also by months to see the different activity of the system along the days of the week or monthly.

As it has been seen, all the analysis will be carried out using the Activity, which is the standard deviation of the number of bikes. In terms of analysis, either the standard deviation of bike's occupation or their average values are worth and meaningful analysing. Nevertheless, since the standard deviation involves X^2, the values are not cancelled out when we are aggregating them in hours, days and months while mean values are cancelled out so their aggregation does not represent changes in that duration.

Furthermore, when a station is empty or reaches its maximum capacity, the mean value does not reflect this inefficiency in the system while the standard deviation becomes small in these situations and represents lower activity. Moreover, the batch transportation of bikes, carried out by Bicing employees, affects more the aggregation of mean values than the aggregation of STD values.

In short, although the analysis of both capacity and activity can provide invaluable information from different perspective, in this case we want to focus and activity and standard deviation is the best proxy to it.

2.5 Principal Component Analysis

2.5.1 General overview

The central idea of principal component analysis (PCA) is to reduce the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and which are ordered so that the first few retain most of the variation present in **all** of the original variables. [7].

Principal Component Analysis attempts to draw straight, explanatory lines through the data and each of these lines represents a principal component (or a relationship between an independent and a dependent variable). While there are as many Principal Components as there are dimensions in the data, PCA's role is to prioritize them. Furthermore, Principal Components are perpendicular.

In Figure 8, the two first Principal Components are depicted. As it can be seen, in blue there is the first principal component whilst in red there is the second principal component.

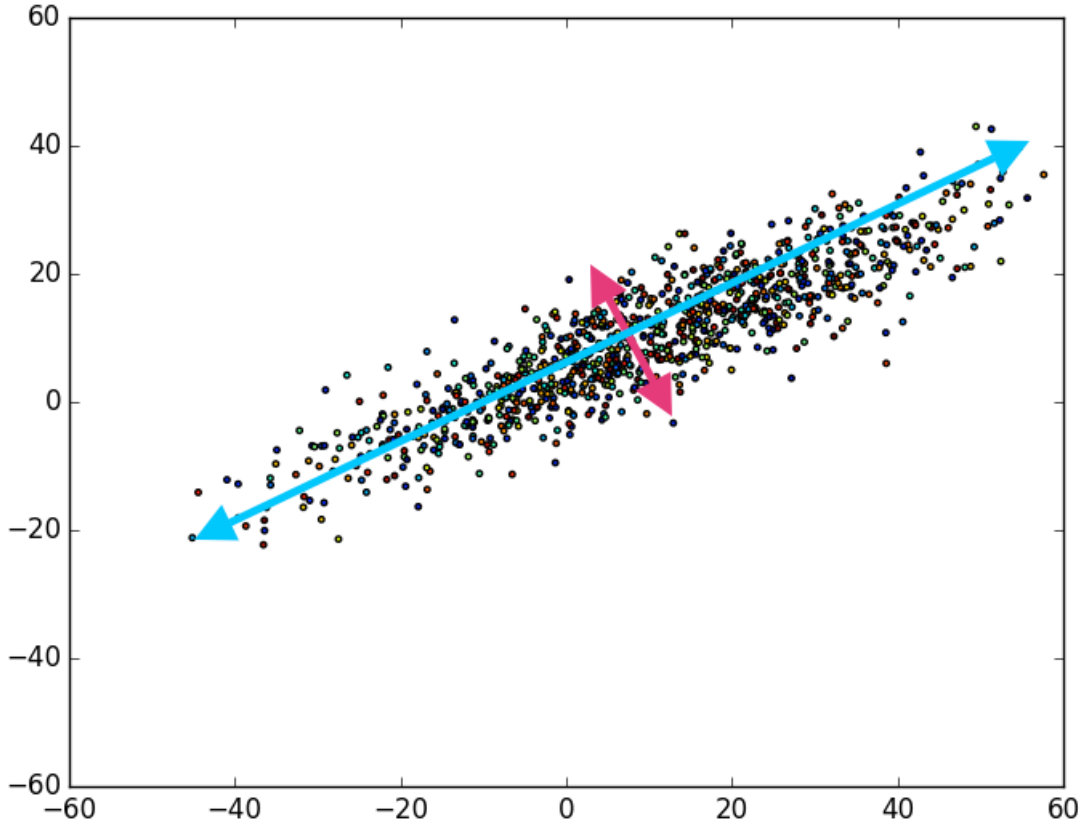

Figure 8: An example of the two first principal components (PC's) of PCA. In blue, 1st PC and 2nd PC in red.

In order to find these principal components, is necessary to calculate the covariance matrix which is $X \cdot X'$, where X is the matrix shown in Figure 7. This matrix is composed in a way that it describes the variance of the data, and the covariance among variables. Finding the eigenvectors and eigenvalues of the covariance matrix is the equivalent of fitting those straight, principal-component

lines to the variance of the data because eigenvectors *trace the principal lines of force*, and the axes of greatest variance and covariance illustrate where the data is most susceptible to change.

Eigenvalues are simply the coefficients attached to eigenvectors, which give the axes magnitude. In this case, they are the measure of the data's covariance. By ranking the eigenvectors in order of their eigenvalues, highest to lowest, we get the principal components in order of significance.

If two variables increase and decrease together, they have a positive covariance, and if one decreases while the other increases, they have a negative covariance.

PCA is particularly advantageous if a set of data with many variables lies, in reality, close to a two-dimensional subspace (plane). In this case the data can be plotted with respect to these two dimensions, thus giving a straightforward visual representation of what the data look like, instead of appearing as a large mass of numbers to be digested. Even with a few more dimensions it is possible, with some degree of ingenuity, to get a "picture" of the data.

2.5.2 Theoretical Background

Principal Component Analysis has either a probabilistic derivation, based on variance maximization and decorrelation of the low-dimensional variables, or a geometrical derivation, also known as minimum reconstruction error. In this case, only the probabilistic derivation is considered [8].

Let us consider we have D observed variables sampled N times. PCA is a linear method, which seeks a linear projection of the high D-dimensional random vector $\boldsymbol{\chi}$ into a lower d-dimensional random vector $\boldsymbol{\Upsilon}$:

$$\boldsymbol{\Upsilon}\ (d \times 1) \quad = \quad \boldsymbol{P}\ (d \times D) \quad \boldsymbol{\chi}\ (D \times 1)$$

In practice, the dimensionality reduction is performed on the basis of sampling points, i.e. finding a lower-dimensional embedding of the high dimensional points as

$$\boldsymbol{Y}\ (d \times N) \quad = \quad \boldsymbol{P}\ (d \times D) \quad \boldsymbol{X}\ (D \times N)$$

Such low-dimensional representation is useful to understand, visualize, classify, etc., high-dimensional data. Furthermore, given a new high-dimensional point $\boldsymbol{x} \in \mathbb{R}^D$, we can readily find its low-dimensional representation as $\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x}$. The reduced variables are commonly referred to as *latent variables*.

Let us write the projection matrix in terms of row vectors

$$\boldsymbol{P} = \begin{pmatrix} -\boldsymbol{p}_1^T - \\ -\boldsymbol{p}_2^T - \\ \vdots \\ -\boldsymbol{p}_d^T - \end{pmatrix},$$

where $\boldsymbol{p}_j$ are $D \times 1$ matrices. Consequently, the reduced representation of a vector $\boldsymbol{x}$ can be written as

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x} = \begin{pmatrix} \boldsymbol{p}_1^T \boldsymbol{x} \\ \boldsymbol{p}_2^T \boldsymbol{x} \\ \vdots \\ \boldsymbol{p}_d^T \boldsymbol{x} \end{pmatrix},$$

where, if the vectors $\{\boldsymbol{p}_1, ..., \boldsymbol{p}_d\}$ are mutually orthogonal, we recognize that these are the components of the orthogonal projection of $\boldsymbol{x}$ onto the subspace spanned by $\{\boldsymbol{p}_1, ..., \boldsymbol{p}_d\}$ in the basis defined by these vectors. To define the first principal component $\boldsymbol{p}_1$, we maximize the variance of the projection of the random vector $\boldsymbol{\chi}$ onto $\boldsymbol{p}_1$, i.e. we want to retain as much variability of the original data as possible in our reduced description. Thus, the goal is to maximize the $\text{Var}[\boldsymbol{p}_1\boldsymbol{\chi}]$, and for convenience we choose the principal component to be a unit vector, i.e. $\boldsymbol{p}_1^T\boldsymbol{p}_1 = 1$. Approximating the variance by the sample variance, and after some manipulations, we have

$$\begin{aligned}\text{Var}\left[\boldsymbol{p}_1\boldsymbol{\chi}\right] = \text{E}\left[(\boldsymbol{p}_1^T\boldsymbol{\chi} - \boldsymbol{p}_1^T\boldsymbol{\mu})^2\right] &\approx \frac{1}{N}\sum_{\alpha=1}^{N}\left(\boldsymbol{p}_1^T\boldsymbol{x}_{\boldsymbol{\alpha}} - \boldsymbol{p}_1^T\bar{\boldsymbol{x}}\right)^2 \\ &= \frac{1}{N}\sum_{\alpha=1}^{N}\left(\boldsymbol{p}_1^T\boldsymbol{x}_{\boldsymbol{\alpha}} - \boldsymbol{p}_1^T\bar{\boldsymbol{x}}\right)\left(\boldsymbol{p}_1^T\boldsymbol{x}_{\boldsymbol{\alpha}} - \boldsymbol{p}_1^T\bar{\boldsymbol{x}}\right)^T \\ &= \frac{1}{N}\sum_{\alpha=1}^{N}\boldsymbol{p}_1^T\left(\boldsymbol{x}_{\boldsymbol{\alpha}} - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_{\boldsymbol{\alpha}} - \bar{\boldsymbol{x}}\right)^T\boldsymbol{p}_1 = \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1.\end{aligned}$$

Consequently, we can find $\boldsymbol{p}_1$ by solving the optimization problem

$$\begin{aligned}\text{Maximize} \quad & \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1 \\ \text{subject to} \quad & 1 - \boldsymbol{p}_1^T\boldsymbol{p}_1 = 0.\end{aligned}$$

In order to solve the optimization problem we introduce the Lagrangian function.

$$\mathcal{L}\left(\boldsymbol{p}_1, \lambda_1\right) = \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1 + \lambda_1\left(1 - \boldsymbol{p}_1^T\boldsymbol{p}_1\right).$$

By making the Lagrangian function stationary, we find necessary conditions for the constrained minimizer, which in principle allow us to solve for the constrained optimizer and the Lagrange multiplier.

$$\boldsymbol{0} = \nabla_{\boldsymbol{p}_1}\mathcal{L} = 2\left(\boldsymbol{S}\boldsymbol{p}_1 - \lambda_1\boldsymbol{p}_1\right) \Longleftrightarrow \boldsymbol{S}\boldsymbol{p}_1 = \lambda_1\boldsymbol{p}_1$$

We can immediately recognize that $\boldsymbol{p}_1$ needs to be an eigenvector of the sample covariance matrix, and λ_1 its associated eigenvalue. Furthermore, the objective function wee seek to maximize is

$$\boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_1 = \lambda_1\boldsymbol{p}_1^T\boldsymbol{p}_1 = \lambda_1,$$

and consequently, $\boldsymbol{p}_1$ should be the eigenvector of the largest eigenvalue.

Let us look for the second principal component $\boldsymbol{p}_2$. We will seek again to maximize the variance of the second component of the reduced description $\text{Var}\left[\boldsymbol{p}_2^T\boldsymbol{\chi}\right] \approx \boldsymbol{p}_2^T\boldsymbol{S}\boldsymbol{p}_2$ subject to normalization of $\boldsymbol{p}_2$ and to decorrelation with the first principal component. Correlation indicates redundancy in the original data, and therefore it is natural to try to eliminate redundancy in the different components of the reduced description. Mathematically, we express decorrelation as

$$0 = \text{Cov}\left[\boldsymbol{p}_1^T\boldsymbol{\chi}, \boldsymbol{p}_2^T\boldsymbol{\chi}\right] \approx \boldsymbol{p}_1^T\boldsymbol{S}\boldsymbol{p}_2 = \lambda_1\boldsymbol{p}_1^T\boldsymbol{p}_2,$$

which thus is an orthogonality condition between $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$. The Lagrangian function is now

$$\mathcal{L}(\boldsymbol{p}_2, \lambda_2, \gamma) = \boldsymbol{p}_2^T \boldsymbol{S} \boldsymbol{p}_2 + \lambda_2 \left(1 - \boldsymbol{p}_2^T \boldsymbol{p}_2\right) + \lambda \boldsymbol{p}_1^T \boldsymbol{p}_2.$$

where γ is the Lagrange multiplier for the orthogonality constraint. The stationarity condition with respect to $\boldsymbol{p}_2$ is now

$$\mathbf{0} = \nabla_{\boldsymbol{p}_2} \mathcal{L} = 2\left(\boldsymbol{S}\boldsymbol{p}_2 - \lambda_1 \boldsymbol{p}_2\right) + \gamma \boldsymbol{p}_1.$$

Pre-multiplying this equation by $\boldsymbol{p}_1^T$ we obtain that $\gamma = 0$, and thus this condition is equivalent to

$$\boldsymbol{S}\boldsymbol{p}_2 = \lambda_2 \boldsymbol{p}_2.$$

As before, the objective function to be maximized is λ_2. Assuming for simplicity that there are no repeated eigenvalues, we cannot take λ_2 to be again the largest eigenvalue (λ_1) because its corresponding eigenvector needs to be orthogonal to $\boldsymbol{p}_1$. Consequently, the second principal component $\boldsymbol{p}_2$ is the eigenvector associated with the second largest eigenvalue of the sample covariance matrix.

We can proceed analogously to find subsequent principal component, in principle D of them. Although in practice we will only consider $d << D$ of them, suppose we have all of them, and therefore $\boldsymbol{P}$ is a $D \times D$ matrix. Defining the diagonal matrix $\boldsymbol{\Lambda}\cdot$ containing the ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D \geq 0$, we can summarize the relation between eigenvectors and eigenvalues as

$$\boldsymbol{S}\boldsymbol{P}^T = \boldsymbol{P}^T \boldsymbol{\Lambda}.$$

Since $\boldsymbol{P}$ is made out of mutually orthogonal unit vectors, it is an orthogonal matrix satisfying $\boldsymbol{P}^{-1} = \boldsymbol{P}^T$. Consequently, we can diagonalize the symmetric semi-positive-definite matrix $\boldsymbol{S}$ as

$$\boldsymbol{\Lambda} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{P}^T.$$

the diagonal matrix $\boldsymbol{\Lambda}$ is the sample covariance matrix of the transformed variables $\boldsymbol{y}$, mutually decorrelated (minimally redundant) and ranked from maximal to minimal variance.

In practice, PCA works well if the spectrum of $\boldsymbol{S}$ decays quickly, i.e. after some d, the eigenvalues λ_i for $i > d$ are very small. The decay of the spectrum provides information about the intrinsic dimensionality of the data. A common practice is to pick d such that

$$0.95 = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{D} \lambda_i},$$

and retain only the first d rows of $\boldsymbol{P}$. We say then that the latent variables bear 95% of the global variance.

3 RESULTS ANALYSIS

After carrying out all the preparation steps described in Section 2, the data has been analysed with two different approaches.

First of all, an exploratory analysis has been done using the standard deviation (proxy for activity). This analysis is divided in three different parts depending on the time-grouping criteria.

On the other hand, a Principal Component Analysis has been performed on the data and it is divided into two different parts, Temporal and Spatial analysis.

3.1 Exploratory analysis results

The aim of the present section is to understand and analyse the results obtained from a statistical point of view. There are three different approaches depending on how the data has been aggregated. These are: Hourly results, Daily results and Monthly results.

3.1.1 Hourly results

Figure 9 consists on a box plot of the activity aggregated hourly. In other words, there are 24 different whiskers that represent all the hours of 2015. Using box plot allows us to see de median values, the different quartiles, the maximum and minimum values of activity and also the outliers. Moreover, the length of the whisker stands for the variability of the activity. So, the longer the whisker, the higher the variability.

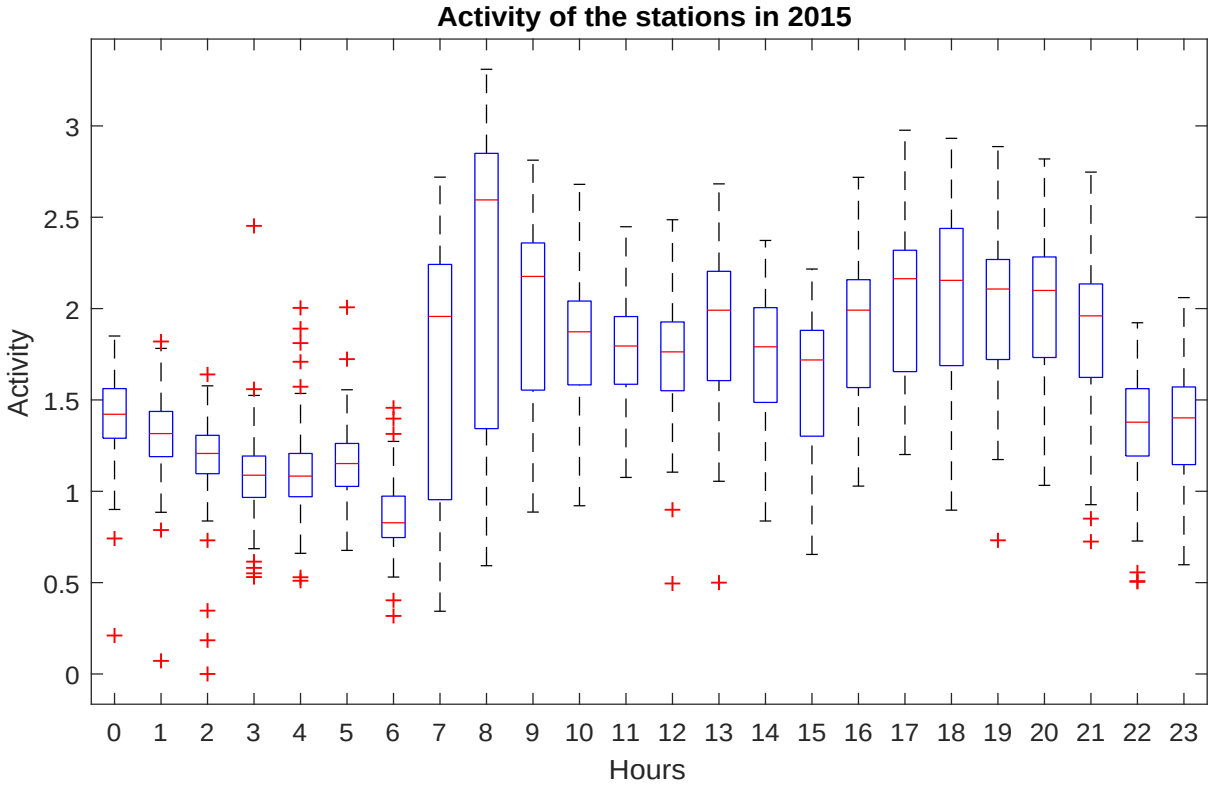

Figure 9: Activity of all the stations in 2015 divided by hours

As aforementioned, Figure 9 depicts the activity of all the stations. The aggregation goes from 00:00 to 00:59 (0), 01:00 to 01:59 (1) and so on. Furthermore, it is useful to mention that Bicing works from Monday to Thursday (05:00 - 02:00) and Friday, Saturday and Sunday (non-stop service).

- **From 00:00 to 06:59**: In general, the activity is low and there is low variability as the whiskers are short. It can be seen that until 02:00 the activity is larger than afterwards because the service is open until that time. Although it is open, the activity is not much larger because during the week people has to work or go to university on the morning. The interesting thing during this time-period are the outliers. These are either larger or smaller than the maximum/minimum values of the distribution and are due to the activity on the weekends and also to the batch transportation of bikes carried out by Bicing employees. Week days may be more predictable but the usage of the bikes during the night at weekends is not. Consequently, the major part of the outliers are located between 00:00 and 06:59.

- **From 07:00 to 09:59**: There is a peak corresponding to Barcelona's citizens going to work, university, school, etc. The most common starting time is between 08:00 and 09:00, hence the period with higher activity is located between this time-period. However, the boxes are really long which means that there is a large variability on the activity during this period. Meteorological conditions, weekends, holidays and other events that break the daily routine can affect the activity of Bicing thus increasing variability.

- **From 10:00 to 12:59**: After the peak at 08:00 there is a decrease of the activity in the system. During this period, most of the users are at work or university and the standard deviation remains steady. However, its levels are higher than the ones at night as not everybody is working. Bikes are being used as citizens are not sleeping but there are no schedules now. In terms of variability, the whiskers are short hence the variability of the data is small.

- **From 13:00 until 15:59**: There is another peak in the activity at 13:00 due to lunch time. Either workers or students finish their shifts/lessons and borrow the bikes to go lunch. Following this peak there is a decrease on the activity while everybody is having lunch. As happened at the first peak of the day, the variability is large. Although it does not reach the same level than before, it has increased with respect to the previous hours. This variability will remain steady until night-time.

- **From 16:00 to 21:59**: Since the schedules of workers and students does not coincide, there is not a common time to leave either work or school. Consequently, the activity is similar during this period as people uses the service gradually. However, note that at 18:00 the box is larger than the rest of the hours thus the data is more variable. This might be due to the aforementioned leaving time which normally starts around that hour.

- **From 22:00 to 23:59**: The activity decays dramatically, reaching night levels. Although the variability is larger than the one at night activity, the usage patterns at that time are similar than the ones at 00:00. The usage of bikes does not follow any work-based pattern and users are mainly using the service to personal use (go play sports, go out, spend free time).

This interpretation has been done based on the results obtained from the statistical analysis of the data for an entire year (2015) and also the intuition I have, as a citizen, of Barcelona and its residents.

3.1.2 Daily results

While data aggregated by hours depicts the trends at different day-times, when the data is grouped by days the results do not unfold any information that we did not have before.

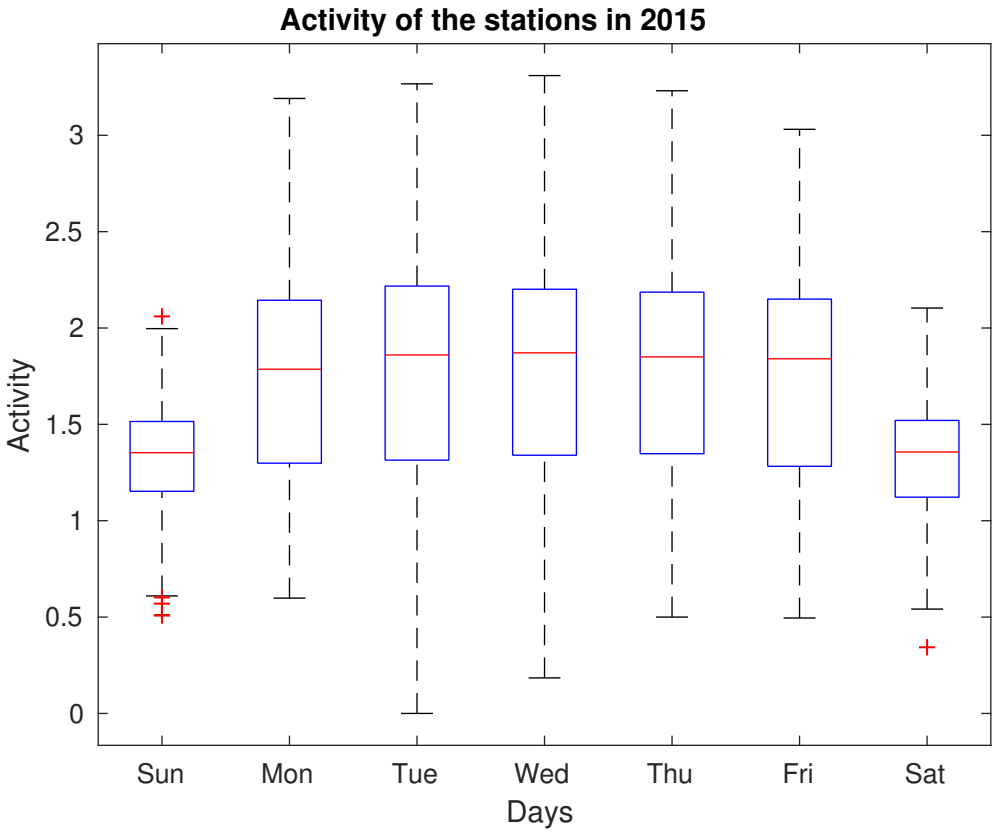

Figure 10: Activity of all the stations in 2015 divided by days

Figure 10 shows the distribution of the activity by days. As it is obvious, that activity is higher on the weekdays than on weekends. Either the median value of the activity or the length of the whiskers is the same from Monday to Friday which means that in general, the activity is equally distributed during these days. Maximum and minimum values are quite away from the median but this is normal because there is no distinction between normal days and holidays or sunny and rainy days.

On the other hand, the weekends have also the same variability and median hence the same activity. It is interesting though the fact that the outliers are located on the weekends, where the activity is more aleatory.

Although the analysis for the data aggregated daily does not reveal any unknown information of the system, it confirms, based on an entire year of data, the fact that weekdays are busier than weekends in terms of Bicing usage.

3.1.3 Monthly results

Before understanding the monthly activity, we need to understand how the weather affects the behaviour of Bicing users. Conditions for using a bike depend on everybody but there are some that may apply to all users. For instance, when it rains the users will not use the bike as it is more dangerous and they will also get wet. Consequently, rain means low activity. The same will happen when it is cold. If temperatures are really low, the users may prefer to use another transport method to avoid getting cold and uncomfortable. On the other hand, sunny days will be optimal for going around with a bike. Nevertheless, too sunny and hot days may have a counter effect, reducing as well the usage of bikes. With the results achieved, let's compare them using weather data from Observatori Fabra [9]. See Table 1 and Figure 11.

Table 1: Average of temperatures and precipitations of 2015 divided by months

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Average T (°C)	9	8.2	12	14.8	19.1	23.4	25.9	23.5	19.7	16.7	14.3	12.6
Precipitation (mm)	13.5	11.3	48	12.5	53.2	10.8	18.2	49	60	29.6	38.7	0.3
Precipitation (%)	4	3	14	4	15	3	5	14	17	9	11	0

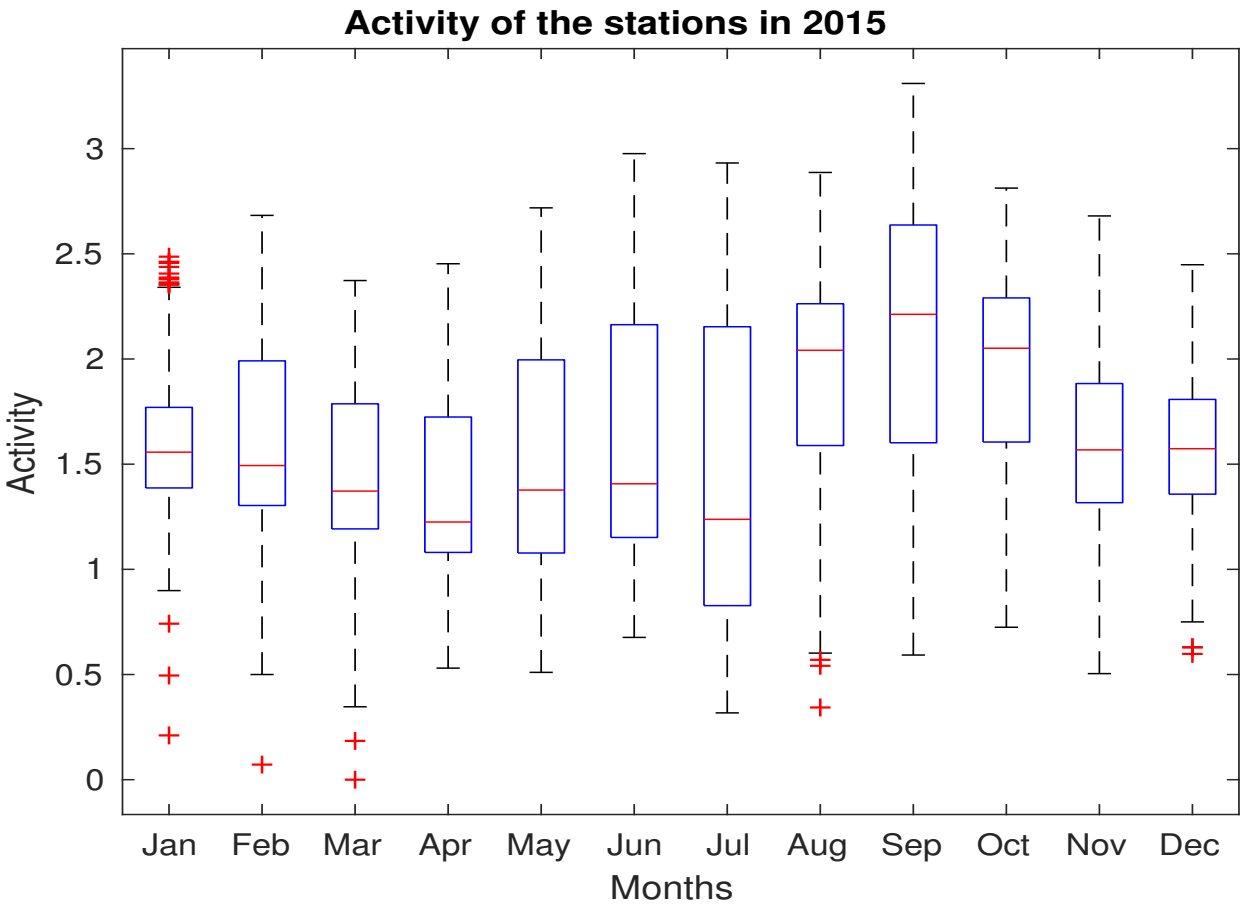

Figure 11: Activity of all the stations in 2015 divided by months

Said that, let's begin examining the monthly activity, starting by **January** and **February**. Both months were by far the coldest months of the year and it did not rain either. January is the first month of the year and normally, everybody start the year with different new "purposes" such as being healthy, using bike instead of private car. This might explain the top outliers on the distribution, as well as some punctual sunny days. On the other hand, the outliers below are caused by the holidays of January or the rainy days as well. Regarding the median value of activity, this is quite similar to February, except that as February was colder, the activity was lower. Furthermore, the variability of the distribution is larger in February than in January. This difference in variability might be explained by lack of holidays in February.

Following our intuition about the weather influence in the bike usage, **March** and **April** activity median value decreased respecting the last two months but the variability remained quite similar than the one in February. Catalan Spring season is characterised by rain and rain means dangerous conditions for cycling. Even though Barcelona is a bike-friendly city with a lot of bike lanes through all the city, it is still a frenetic city with lots of traffic and congestions. If it rains, the traffic gets worse and this increases the danger for cyclists. Consequently, it is normal that the median value decreases on this conditions.

Similarly to March and April, we found **May**. May is still a Spring month and as it is normal, it rains. However, the weather conditions and temperatures are more friendly with less windy days or very cold ones. If we take a look at Table 1, we can see that the Average Temperature in May was no less than 19.1°C. That is a good weather for cycling as it is neither too hot nor to cold. With the arrival of sunny days, it also increases the variability of the distribution.

Following this pattern we see that **June** has also a higher median and variability. It seems reasonable that the activity in June has to increase with respect to previous months. However, the increase of activity is also counter affected by the end of university and school years. Thus the increment of usage due to good weather gets cancelled by the academic year ending.

July 2015 was, according to Observatori Fabra [9], one of the hottest July's since 1914. Maximum temperatures of 36°C make very difficult to stand under the sun at certain hours. This, adding the lack of rain, is translated into a decrease of activity. However, the variability is really large in this month. Holidays and tourists may also affect the bikes activity, using them at reasonable daytimes such as early in the morning, late afternoon or evenings.

If July was the bad cop, **August** was in this case the good one. The average temperature decreased in 2.4°C, reaching a value of 23.5°C. Moreover, it rained a lot (considering the typical dry summers we are used to) and that created better conditions than July. It can be seen as well that the variability of the distribution was shorter and there were also three outliers (which probably correspond to the three days it rained in that month). The high activity might be due to tourists who subscribe only for one or two months to Bicing service (which is economical than renting a personal bike during all this period) and also to locals who enjoy going by bike around when they go out, meet some friends, got to the beach, etc.

September's activity is the highest of all the year. At the beginning of the month, most people are still in holiday, there are also a lot of tourists by that time of the year and moreover, the weather is also warm and friendly. However, it needs to be said that in the dataset, there were some missing days of information that were not available for analysis or only had 2 or 3 hours of information. Those days were removed and they coincide with the second half of September. It seems reasonable that by the second half of the month, the students and workers are back to their normal schedules

and the activity should be reduced (in comparison with holiday activity). But there is no data for that period, hence September is the most active month based on what we have and always taking into account the possible bias due to missing data.

The activity registered on **October** is also surprising. It was a dry October in comparison to other years, and the temperatures where similar (even higher) to the ones in April. The return from holidays, which means our energy is full, and also the good weather in Barcelona helped users to use the bike during that month.

November, as the temperatures decreased and the rains increased, suffered a decay in the activity, reaching levels similar to January or February. And so did **December** which was even cooler but the activity remained steady respecting November. The only difference between these two months was the variability of the distribution, being larger in November as the temperatures were not as cold as December.

3.2 Principal Component Analysis results

In the present section, the different results for the Principal Component Analysis are interpreted. Before starting, it is necessary to explain the two different approaches followed to get the following results. These are a Temporal and a Spatial Analysis (which depend on the covariance matrix we build).

3.2.1 Temporal Analysis

Carrying out a Temporal Analysis consists on considering the time as the variable, and in this case, our variables are the hours of a day. Consequently, the dimension of the Covariance matrix is 24×24 and the Principal Components are vectors whose dimension is equal to 1×24.

In this section, the results for the Temporal PCA approach are shown in Figures 12, 13, 14 and also Table 2. Moreover, these will be interpreted and discussed.

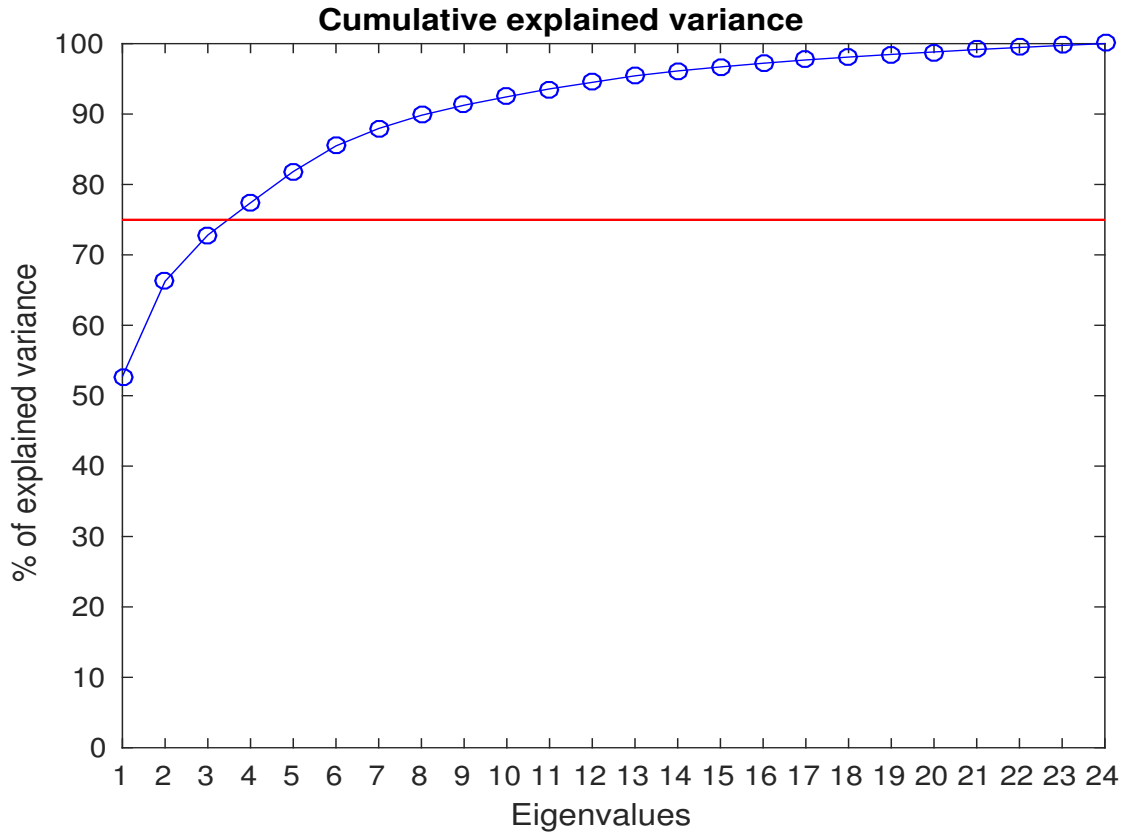

Figure 12: Cumulative sum of the explained variance by the principal components for the temporal analysis

Figure 12 depicts the cumulative sum of the explained variance. The number of dots in the graph is equal to the number of eigenvalues (24, corresponding with the variables). It is interesting to see that the first principal component explains the 53% of the total variability of the data. If we add the second principal component, the explained variability is more or less 66%. And with only three principal components we are able to explain almost 75% of the data.

Taking that into account, it is possible to have more than a general idea about the behaviour of the Bicing users with only using three dimensions instead of 24 which was the initial dimension value of this problem. Reducing dimensionality of the data is particularly interesting when dealing with Big Data problems. This reduction can reduce storage and computational time, achieving also accurate results.

Before starting analysing what represents each principal component, let's examine Table 2 which shows the three first PC's for the temporal analysis. To gain visibility on the existing correlation, the values have been transformed to signs according the following criteria:

- $x \geqslant 0.25 \Longrightarrow (+)$ **&** $x \leqslant -0.25 \Longrightarrow (-)$
- $0.1 \leqslant x < 0.25 \Longrightarrow +$ **&** $-0.25 < x \leqslant -0.1 \Longrightarrow -$
- $x < 0.1 \Longrightarrow$ *no sing* **&** $x > -0.1 \Longrightarrow$ *no sign*

Now, using Table 2 and Figure 13, we will label the three principal components according to what they explain and they are to be analysed in depth.

Table 2: Three first Principal Components for the temporal analysis

Time	PC1	PC2	PC3
00:00		(+)	
01:00		(+)	
02:00		(+)	
03:00		(+)	
04:00		(+)	
05:00	+	(+)	
06:00			
07:00	+	-	(+)
08:00	(+)		(+)
09:00	(+)		+
10:00	(+)		(+)
11:00	+		+
12:00	+		+
13:00	+		
14:00	+		
15:00	+	-	-
16:00	+		(-)
17:00	(+)		
18:00	(+)		(-)
19:00	+		(-)
20:00	+		(-)
21:00	+		-
22:00	+		-
23:00	+	+	-
Explained Variance	52.7%	66.3%	72.8%

The first principal component (PC1 in Table 2) has its highest values located on the hours with more activity, which are the ones with the (+) sign, according with the explained in the Explanatory Analysis section. Moreover, the other positive signs (which have values between 0.1 and 0.25) represent the "calm periods" between peaks. Depicted in Figure 13 we can see in blue the first principal component. As it stands for the hours with larger demand, from now on this component will be labelled as "High-Demand".

One interesting thing is that all the values of the first principal component are positive. So, in one way or another, exists a correlation between all the hours of the day (as there is activity always). However, taking into account that the largest values are the ones located at peaks and throughout the day, it seems reasonable to relabel this component as "High-Demand".

Obviously, as this component stands for demand, the largest difference occurs between day and night time. More precisely, between night and the morning peak as showed in Table 2 and Figure 13

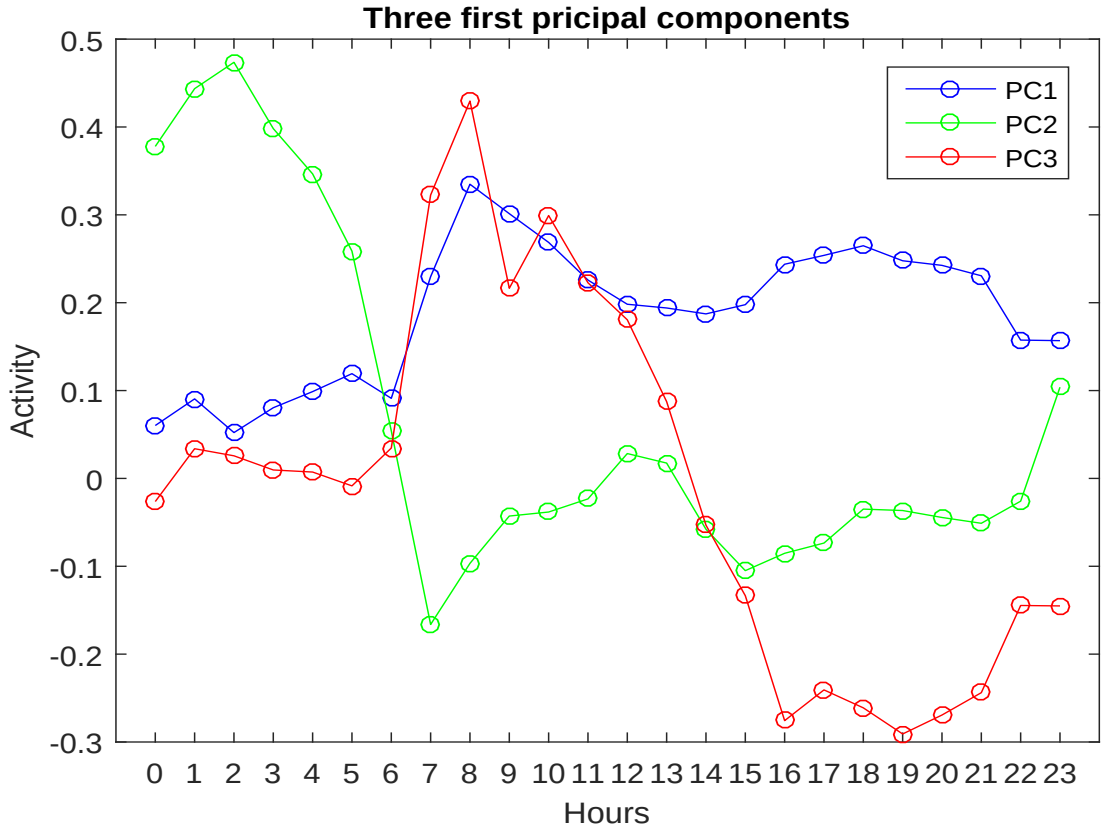

Figure 13: Three first principal components for the temporal analysis

While the first principal component ("High-Demand") explains the high activity hours, the second principal component (depicted in green in Figure 13) seems to explain the night activity (or the low activity hours). Taking into account that, it seems convenient to label this component as "Night vs. Day".

The peaks depicted in Figure 13 correspond with the hours where the demand is low such as night time or during the morning when everybody is at work/school. The off-peaks, or the smallest values are the ones that showed higher activity in "High-Demand" component.

In this case, there are positive and negative values. That is, that the night hours are negatively correlated with the day hours. For instance, this might mean that stations with lower activity during the night have a frenetic activity at day time or the other way around, being the ones with less activity during the day the more active during the night.

If we compare both principal components in Figure 13 it is easy to see that they are more or less opposite, hence it seems appropriate to have relabelled them that way.

Later on this section, the projection of both first and second principal components are plotted and the relationship between these two components is to be interpreted.

We have already seen that the two first principal components represent "High-Demand" and "Night vs. Day" and explain 66% of the data. On the other hand, the third principal component (now depicted in red in Figure 13) explains the activity in the morning and the one in the afternoon. In other words, the activity during the morning is negatively correlated with the activity occurring in the afternoon. This negative correlation is really interesting as we know beforehand that there are stations located uptown from where users will take bikes in the morning but will not bring them back at the afternoon.

Our third principal component explains exactly this pattern. Consequently, it seems interesting to label this component as "Morning vs. Afternoon".

In short, with the three first principal components, labelled as "High-Demand", "Night vs. Day" and "Morning vs. Afternoon" we are explaining 73% of the variability of the data. In other words, it is possible to understand what happens during the whole day with only three components. It is indeed obvious that some information is not being taken into account as there are 24 principal components but with only three we have been able to understand what happens during the day, at night time and between morning and afternoon. Moreover, combining both explanatory analysis and PCA we have gained a general idea (based on data) of the behaviour of the system. Considering the size of the data, we have been able to do this with few plots, so once again, it is worth to remark the importance of the data preparation and the analysis.

Continuing with the analysis of the first principal component, another interesting result is depicted in Figure 14 which is the projection of the first principal component. This projection shows the value of the first principal component over the 419 Bicing stations distributed over the city of Barcelona (we plot them using their coordinates). As aforementioned, the first principal component is the one explaining the high demand of bikes. Consequently, when we project it over the stations, we are seeing in red the stations where the demand is higher and in blue the ones with lover demand.

It is not surprising that the city centre and the coast area are the zones where the activity is higher since the conditions for cycling are better. On the other hand, the up-town areas such as Horta - Guinardó or Sant Andreu appear rather in blue than red showing that the activity there is lower than in other areas. Nevertheless, Sarrià - Sant Gervasi appears in Cyan/Green meaning that the activity is higher than other up-town areas but still low compared with the city centre.

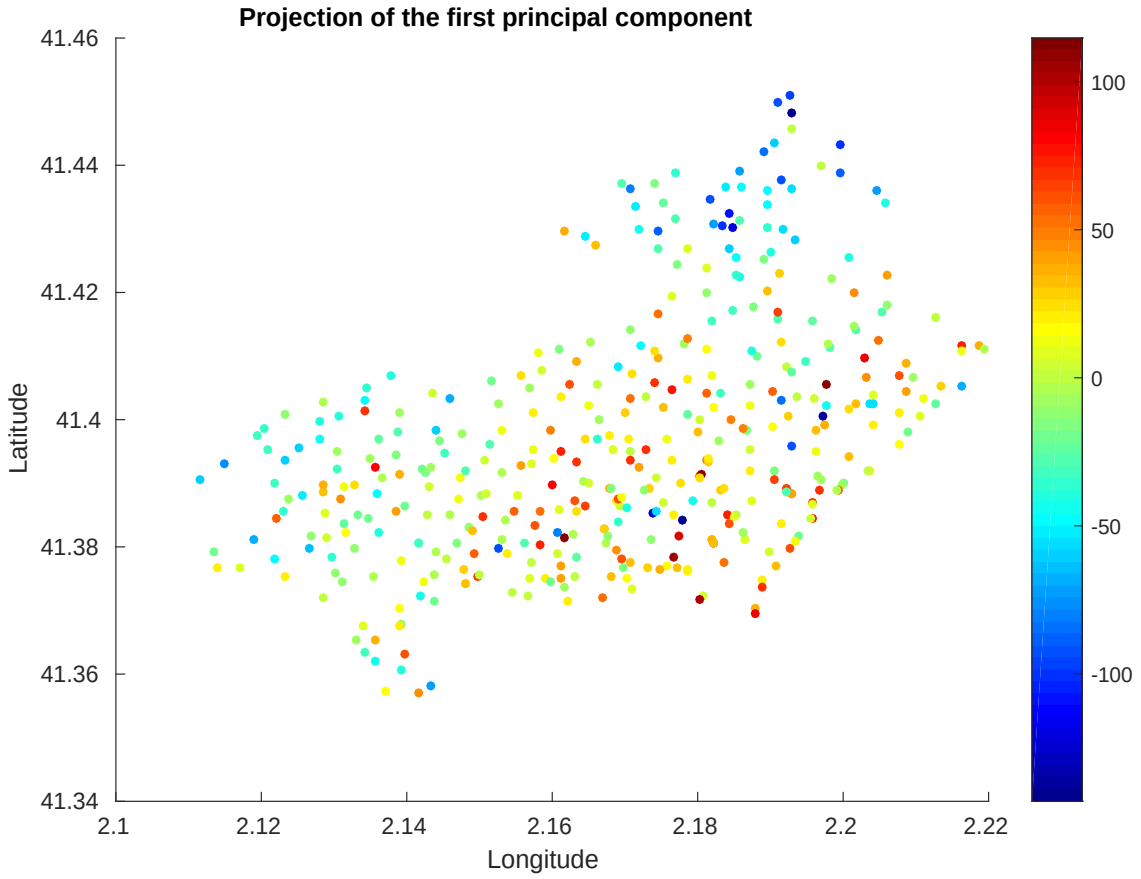

Figure 14: Projection of the First principal component

To sum up, using only the first principal component projected over the stations makes possible to have a general idea about which stations are busier than others and where the activity is concentrated. However, it needs to be remarked that the first principal component explains the 53% of the variability but still is a good result in terms of gaining visibility.

To further understand the behaviour of the system, it is interesting to plot again the projection of the data but this time over the principal components. Figure 14 showed the projection of the first principal component over the stations location. On the other hand, Figures 15, 19 and 22 depict the projections over the first principal component axes. Moreover, some stations with high (or low) values on these plots have been chosen and analysed in depth. Now that we know what each principal component stands for, let's see how the stations are related to these components and prove that principal component analysis is able to describe the real behaviour of the selected stations.

3.2.1.1 First and Second Principal Components

Figure 15 shows the data projected over the first and second principal components, that are "High-Demand" and "Night vs. Day". In this case, each dot represents a single Bicing station and its location on the graph depends on its behaviour. Particularly, the dots located to the right (the ones with higher value on PC1) are the stations where the activity is higher while the ones located on the left are those stations with less activity. On the other hand, we see that the stations positioned on the upper part of Figure 15 are the ones where the activity is higher during the night while the ones at the bottom are the stations with no activity at night.

According to this, it can be seen that the major part of the stations are depicted at the bottom of the graph, showing that, in general, the activity is higher at daytime. Furthermore, it is worth to remark that PC1 explains 53% of the data while PC2 explains only 13%. This fact is now interesting because it means that variations on the horizontal axis (PC1) have a greater impact than variations in the vertical axis (PC2) due to the variability they explain.

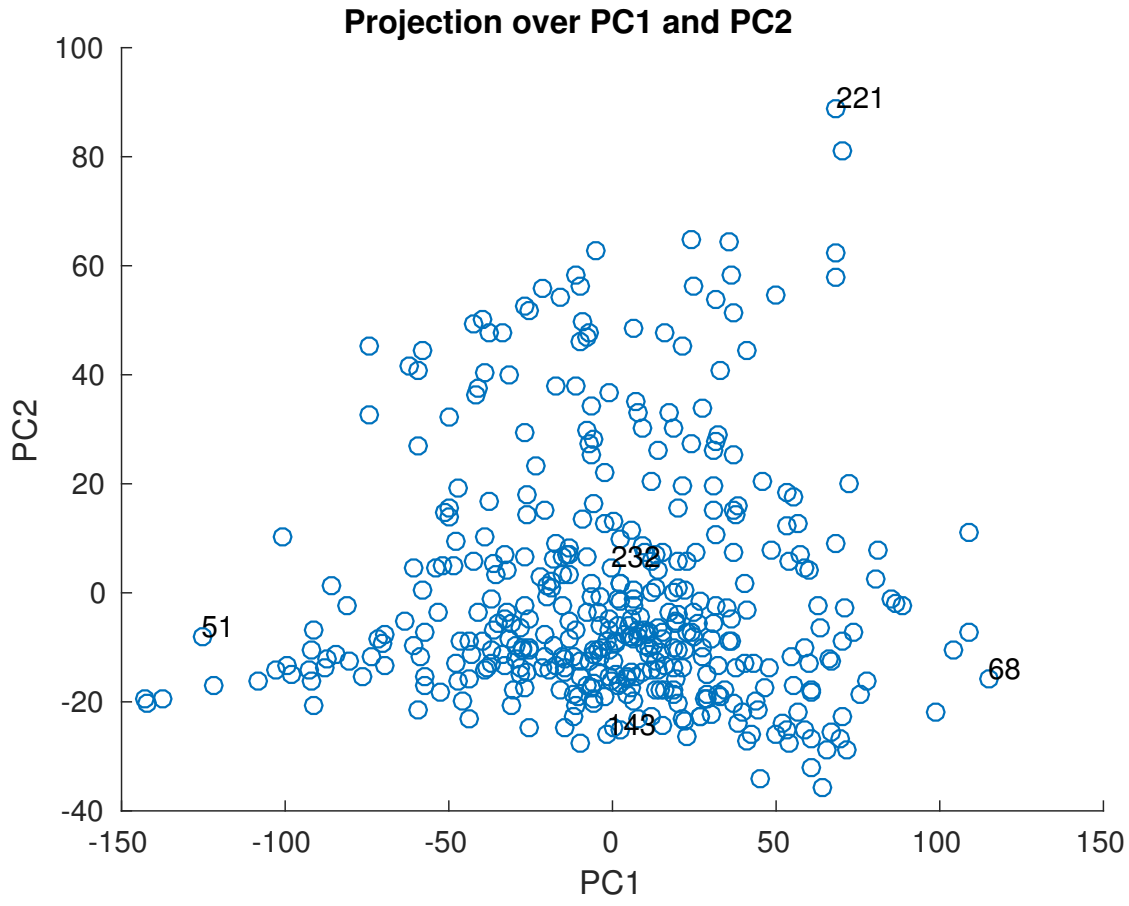

Figure 15: Projection of PC1 vs. PC2 including stations of interest

Let's analyse now the selected stations, which are Station 51, 68, 143, 221 and 232.

- **Station 51 - MACBA**: This station is located on the left side at Figure 15, which according to what we have detailed previously, means that is a station that does not have a large amount of activity, more precisely, during the high-demand hours. To prove that the principal components represent with accuracy what happens in the reality, the activity for Station 51 is depicted at Figure 16 in blue. Station 51 is located at **MACBA**, close to the Faculty of History of the UB. Even though it exists activity during the day, it is higher at the afternoon, when the classes finish and the peaks are located during the night which can be due to batch move of bikes carried out by the employees of Bicing. Even though this stations is at the city centre, the connection with metro and train is really good, affecting then the activity of Bicing on this station.

- **Station 68 - Hotel ARS**: The present station is located on the right side at Figure 15. In this case, the right side represents the part where the First Principal Component has larger values, hence being the activity on the station large during the High-Demand hours. Real activity for Station 68 is depicted at Figure 16. As it can be seen, the activity is really large during the day, as this station is located next to the "Hotel ARS - La Barceloneta" which is a touristic area, next to the sea and bike-friendly (no cars, flat surfaces). Moreover, users who want to go to the beach will leave the bikes in such station, mainly for its location, which explains the high activity during the day. It is interesting to compare both activities to see the difference on activity (hence the different in position at Figure 16).

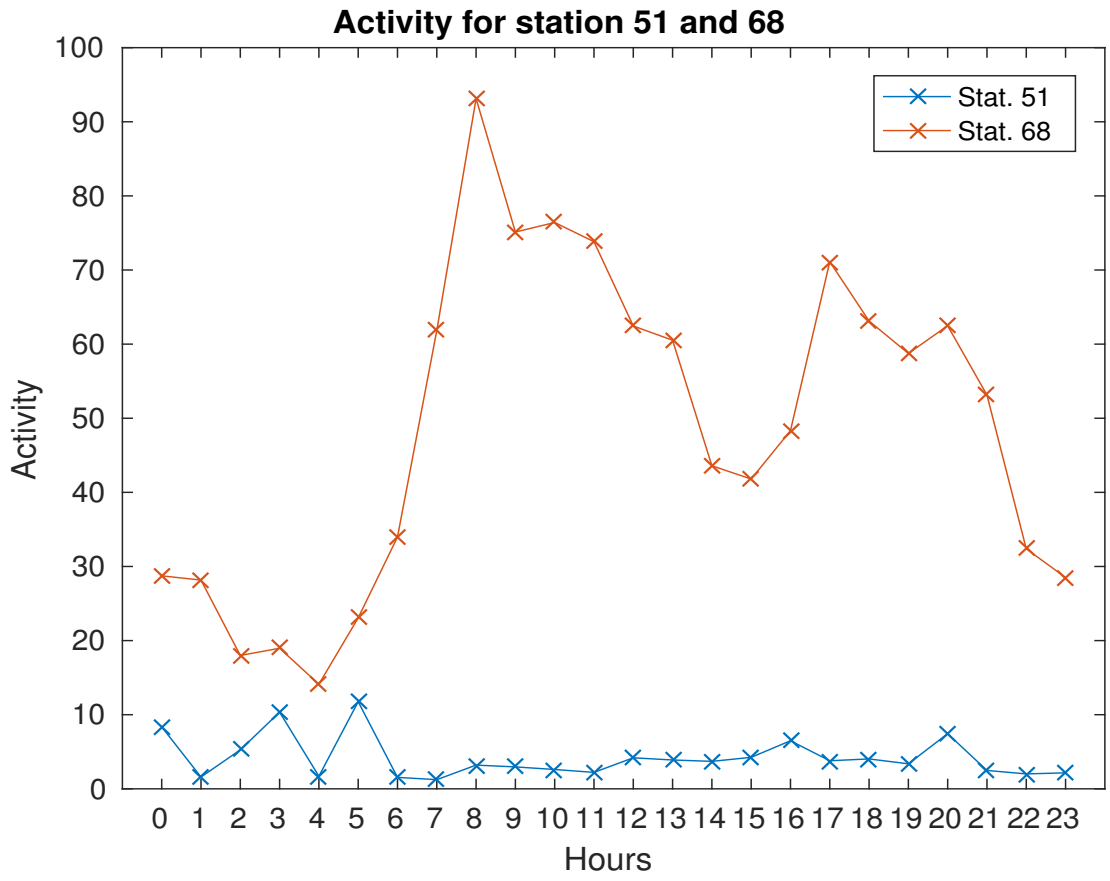

Figure 16: Activity at Stations 51 and 68

- **Station 143 - Diagonal Mar**: Since this station is represented in the lower part of Figure 15 and close to zero for PC1, this can only mean that the station activity is not really high and it has a diurnal behaviour rather than nocturnal. To prove if this behaviour is well explained for our Principal Components, depicted in blue at Figure 17 we find the activity of this station. We can see that the activity has a couple of peaks both during the day and has almost no activity during the night, meeting the Principal Components explanation. This station location is close to offices and commercial centres which explain the diurnal behaviour with peaks at the rush hours and the lack of activity at night.

- **Station 221 - Gràcia amb Diagonal**: Opposite to what happened in Station 143, Station 221 is located un the upper part at Figure 15 and a bit to the right than the origin. This deviation to the right side will be translated into high activity during the day. Moreover, the fact that the Station is located on the upper part will mean that the activity is larger at night than during the day. If we take a look at Figure 17 we can see in red the activity for Station 221. In general, the activity is larger than the previous station since the position is more to the right than before. Nevertheless, the activity has its largest peak during night which is explained due to the location of the Station, close to Pubs and Clubs.

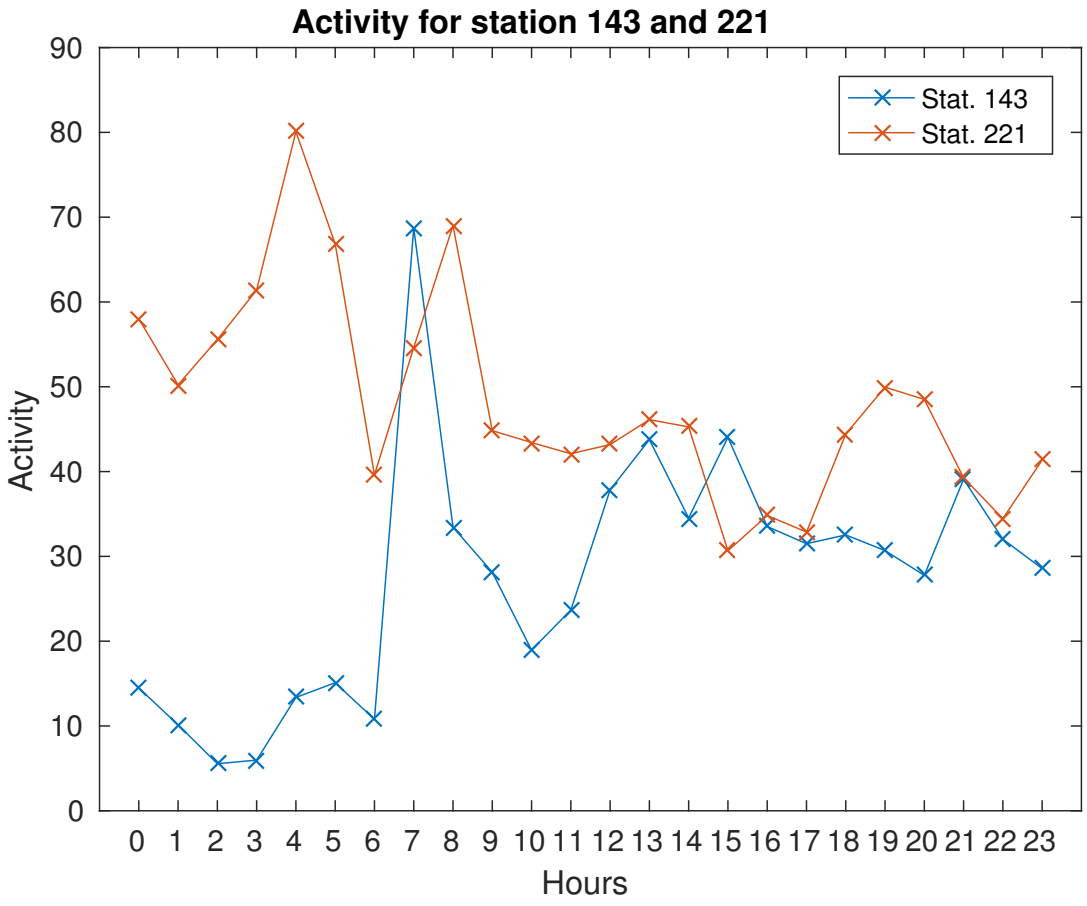

Figure 17: Activity at Stations 143 and 221

- **Station 232 - Montjuïc / Poble Sec**: This station is located in the middle of Figure 15, around the origin (0,0). In terms of analysis, this can only mean that the activity is neither high nor low and the period of the day where the activity is higher is neither the night nor the day. It seems reasonable then to say that this station might be better explained by another Principal Component (for instance PC3, which explains Morning vs. Afternoon activity). In this case, if we check Figure 18 we see that there is a huge peak during the morning and the activity during the rest fo the day is rather low.

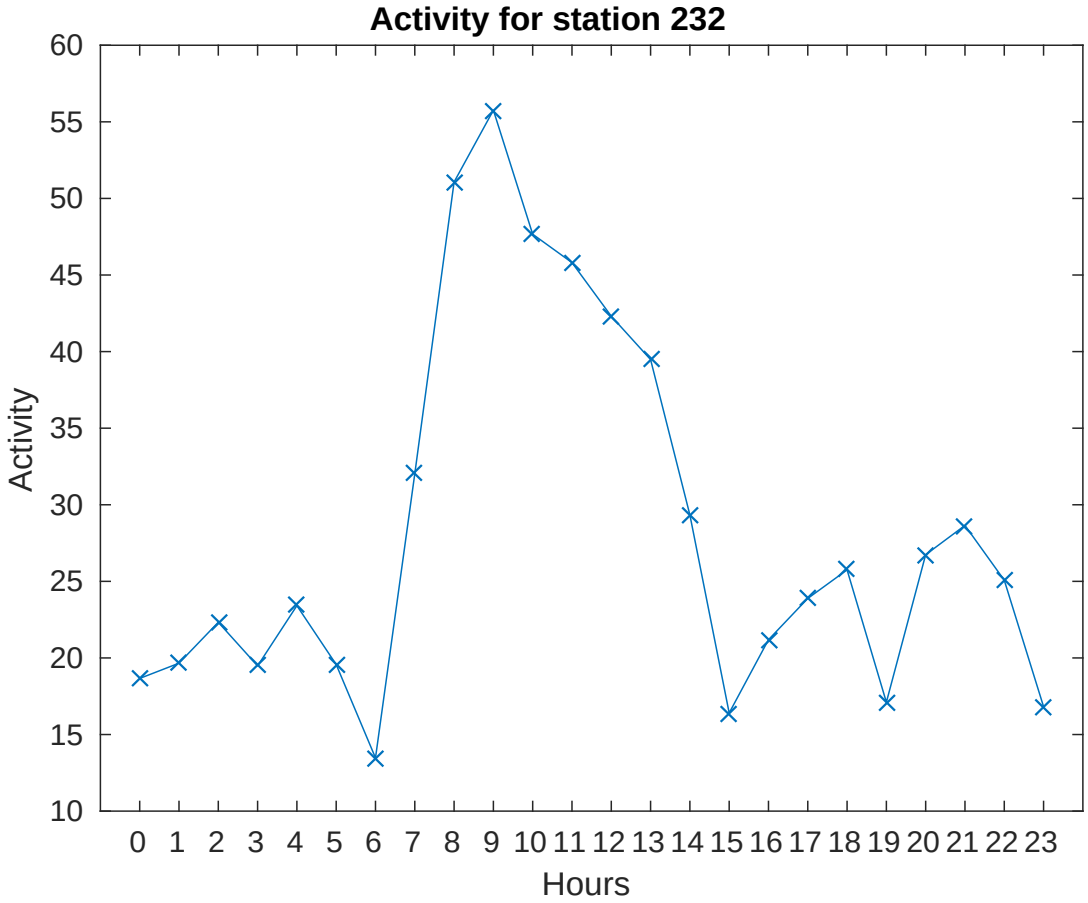

Figure 18: Activity at Station 232

3.2.1.2 First and Third Principal Components

Another interesting way to visualize the principal components is to carry out the same plot but now changing the second principal component for the third one (see Figure 19). Now, the horizontal distribution of dots (or stations) remains constant as PC1 has been used again. Nevertheless, the vertical distribution is different. As depicted, the general trend is that almost all the stations are positioned in the central part of PC3. This component explains the "Morning vs. Afternoon" activity and not all the stations have this exclusive behaviour (which is activity in just one of the periods). So, it makes sense that the majority of the stations are located on the centre being the ones on the extremes the stations with activity either during the morning or the afternoon. The upper part of the graph has the stations with more activity during the morning and the lower part the ones with activity on the afternoon.

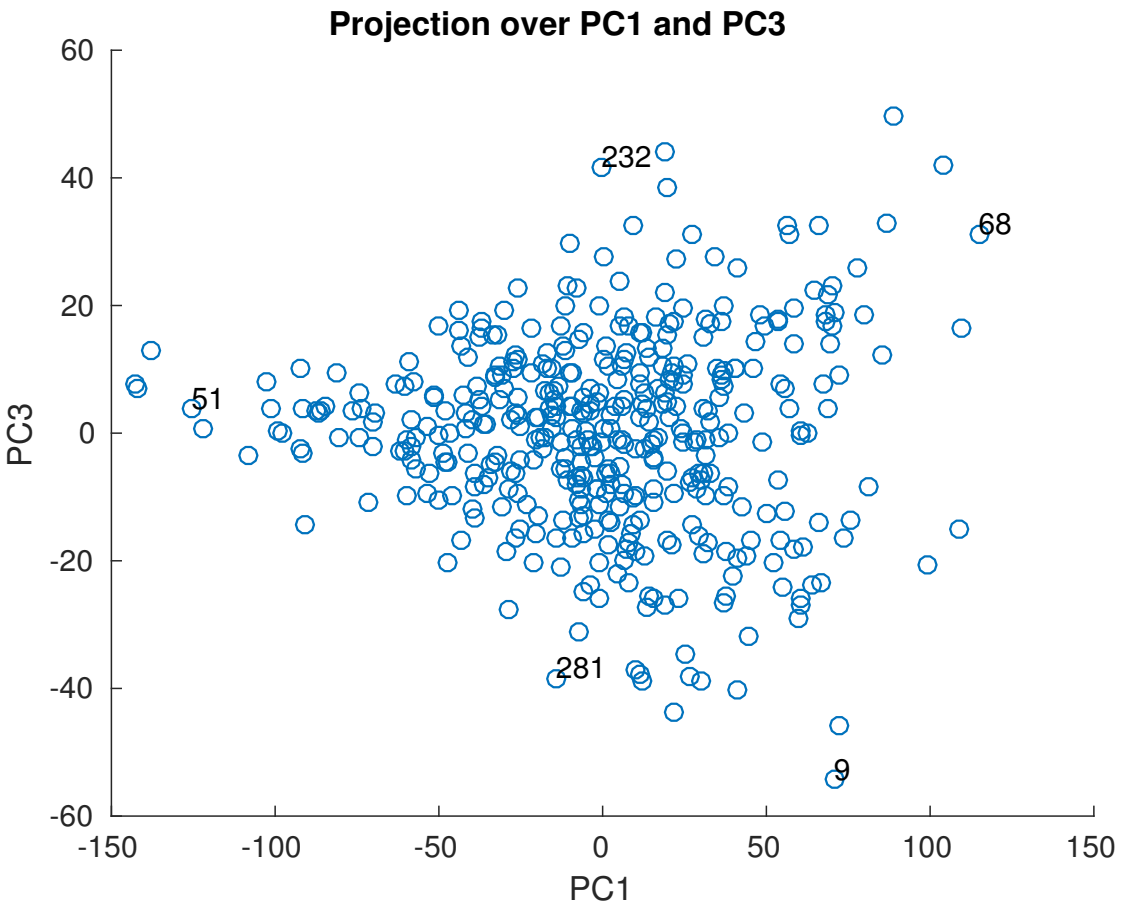

Figure 19: Projection of PC1 vs. PC3 including stations of interest

In order to prove that these two Principal Components also provide reliable results, some stations located in the extremes have been selected as well and are to be analysed. In this case, some of the stations are the ones explained before which appear again on extremal positions. We have tried to choose the same stations once again so it is possible to see the different features each Principal Component explains. And, in order to see different behaviours, there are new stations that are introduced and interpreted.

- **Stations 51 and 68**: Both stations had been previously analysed and we had found that while Station 51 had low activity, Station 68 had a large amount of activity during the day. This was due to the location regarding the First Principal Component (PC1). On the other hand, since PC3 explains the behaviour of the morning or the afternoon, it makes sense that Station 51 is located around 0. Furthermore, Station 68 is located on the right-upper part of Figure 19 which can only mean that it has high activity and also that this activity is larger on the morning that on the afternoon. If we check again Figure 16 we can see that effectively the activity is higher on the morning for Station 68 while there is almost no activity for Station 51.

- **Station 9 - El Born / Ciutadella**: If we take a look at Figure 19 we see that Station 9 is located at the very bottom, on the right side of the graph. Taking into account what we have explained for PC1 and PC2, we know that the activity of this station will be large as its projection has a high value on PC1 axis. Moreover, its position on the vertical axis, which is the lowest we see, means that the activity will be much higher during the afternoon, in comparison with morning activity. If we check now the activity depicted at Figure 20 for Station 9, we see that the behaviour explained by the Principal Components was right. Moreover, we can see that there is also a huge activity during the night, which presumably is explained by PC2 (and will be addressed later on).

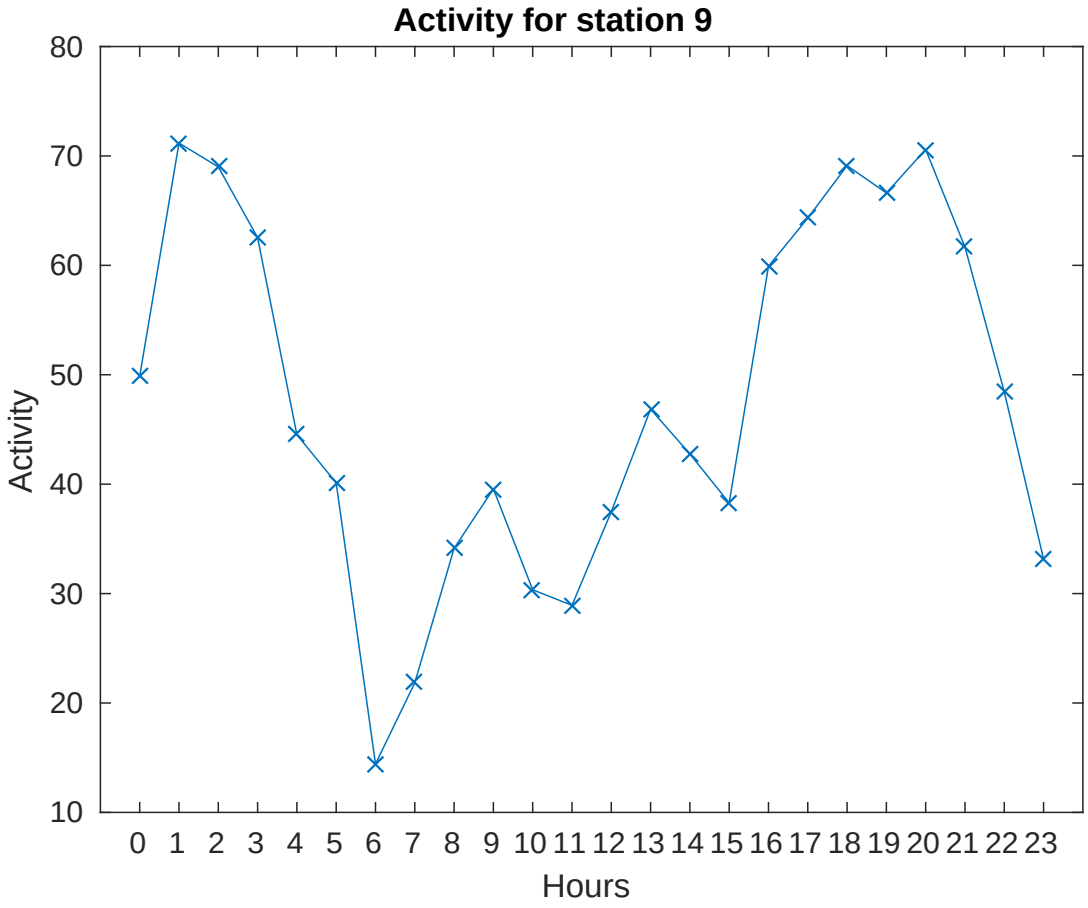

Figure 20: Activity at Station 9

- **Station 232 - Montjuïc / Poble Sec**: This station has been previously explained according to its projection over the PC1 and PC2. Now, we are using PC1 and PC3. AS before, the activity for this location is neither high nor low as it is located around the origin for PC1. On the other hand, this station occupies position on the upper-part of Figure 19 which now is the PC3. Consequently, this station is more active during the mornings. If we take a look now at Figure 21 we can see that the aforementioned behaviour corresponds with the real activity of the station, founding the activity peak on the morning.

- **Station 281 - L'Illa Diagonal**: Oppositely with Station 232, Station 281 is located on the lower-part of Figure 19 which means that the activity is higher on the afternoons. Furthermore, this station is also located around the zero value for PC1 which means that the activity is moderate. Observing Figure 21 it is possible to discover that the general activity is lower than at Station 232 which makes sense according to the horizontal position and that the activity is larger at the afternoons which also fits with its position on the lower part of Figure 21. This peaks at the afternoon are basically explained by the station location on the map. It is uptown and close to offices and universities. Consequently, the activity is higher when the shifts finish.

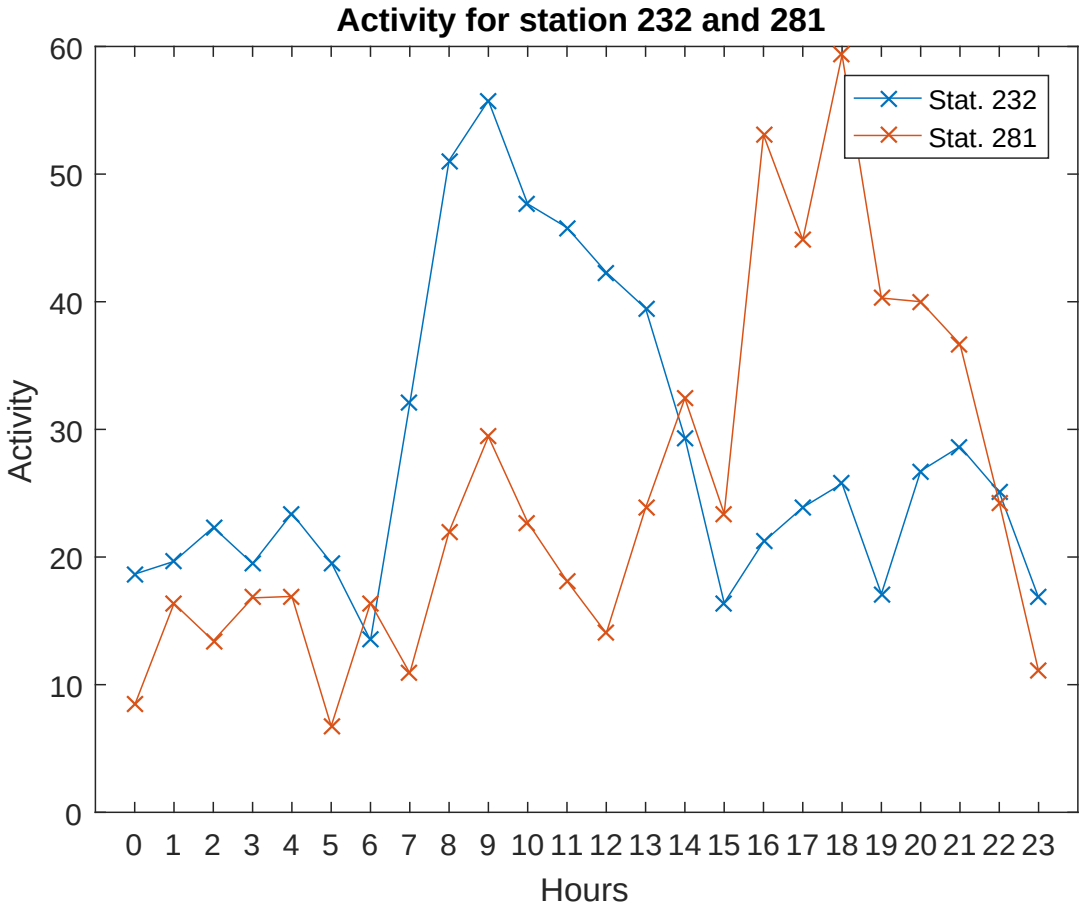

Figure 21: Activity at Stations 232 and 281

3.2.1.3 Second and Third Principal Components

Last but not least, Figure 22 shows the projection over the second and third principal components. The stations that will be analysed have been also included in this Figure.

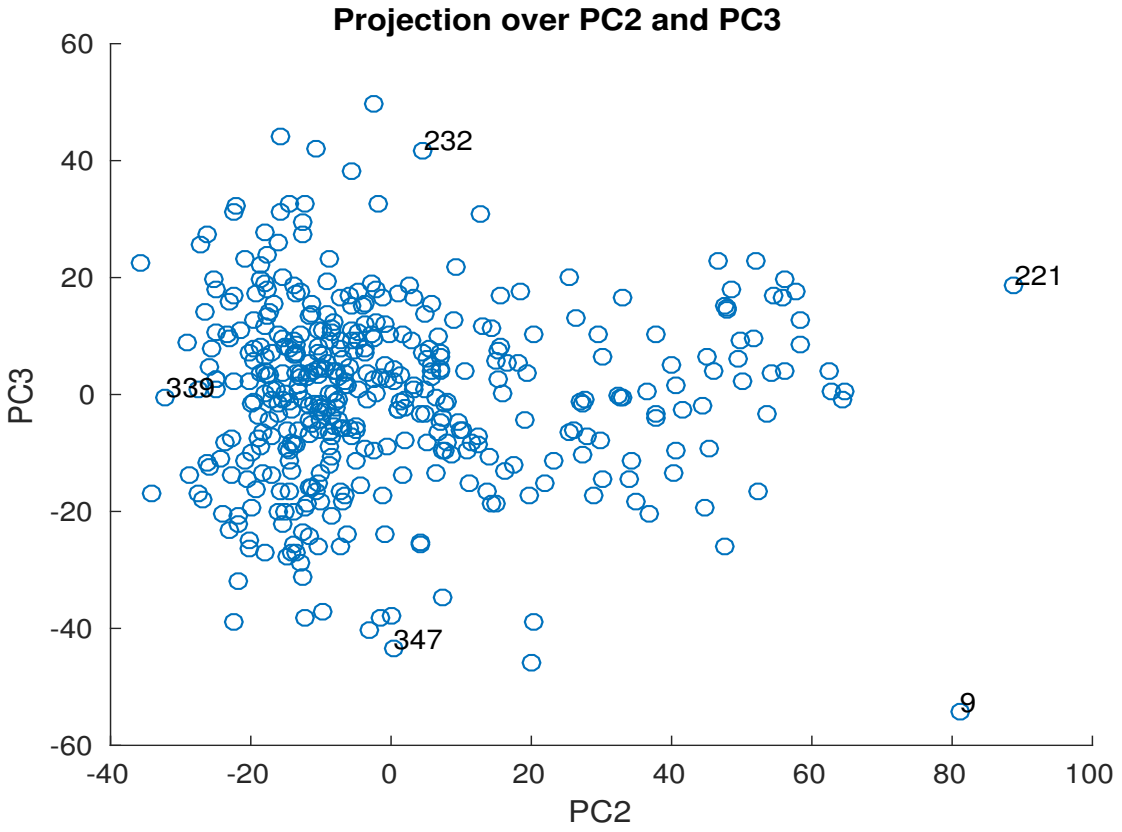

Figure 22: Projection of PC2 vs. PC3 including stations of interest

This time, the vast majority of stations are positioned on the left side of Figure 22 (horizontal axis is the PC2 (or "Night vs. Day")) which is the same as saying that they have diurnal activity rather than nocturnal. Vertical axis, which is the PC3 (or "Morning vs. Afternoon") has a concentration of stations at the centre. As before, not all the stations only have activity during the morning or the afternoon, they normally have activity both periods. The extremes here are those stations which show that behaviour.

The stations that are to be analysed are Stations 9, 221, 232, 247 and 339.

- **Station 9 - El Born / Ciutadella**: Previously, we saw at Figure 20 that this station presented two different peaks, one in the afternoon (explained by the Third Principal Component) and one at night, presumably explained by the Second Principal Component. Now, looking at figure 22 we see that Station 9 is located on the right side of the plot which is the same as saying that the projection over PC2 is really high. Consequently, the activity at night is higher than the activity during the day. In short, for Station 9 we have been able to explain its behaviour using only the three first Principal Components, which have explained the level of activity, the off-peak on the morning, the peak on the afternoon and the peak on the evening.

- **Station 221 - Gràcia amb Diagonal**: We had previously seen that according to PC1, this station is active during the day. Moreover, if we observe now Figure 22, where the stations are represented over the Second and Third PC's, we see that Station 221 is located now on the right side of the Figure and over the vertical origin. Consequently, Station 221 is an active station during the night and also during the morning, being the hours with inferior amount of activity the ones on the afternoons.

- **Station 339 - Glòries**: On the other hand, Station 339 is located on the left and on the vertical origin of coordinates (see Figure 22). In other words, this station is clearly a station with activity during the day rather than during the night. Moreover, there is no difference with the activity during the morning or the afternoon. To check if this fits the reality, we can take a look at Figure 23 and in red we can see that Station 339 has a similar activity during the day hours and a lack of activity at night.

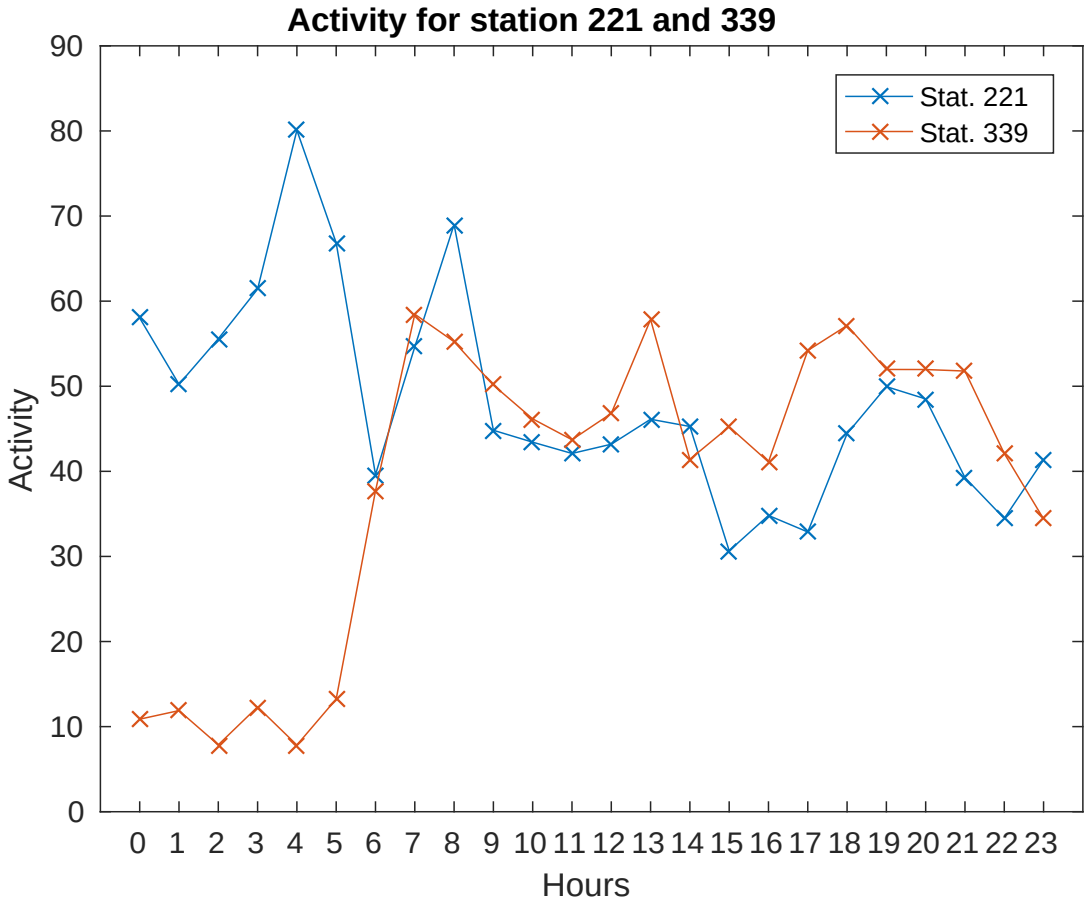

Figure 23: Activity at Stations 221 and 339

- **Station 232 - Montjuïc / Poble Sec**: In the previous two projections we have seen that Station 232 had values close to zero for PC1 and PC2. Oppositely, it has a large value when projected over the Third Principal Component. This can only mean that Station 232 activity happens basically during the morning. If we check again this behaviour explained by the PC3 and compare it with the real activity at the station, we see that they match, being the night and afternoon activity very low in comparison with the morning activity.

- **Station 347 - Fsc. Macià**: The last station that will be analysed is Station 347, which is located in Francesc Macià. At Figure 22 we find this station at the bottom of the plot and once again around the zero for the PC2. To interpret its position is necessary to recall that stations with negative value when projected over PC3 will have be more active on the afternoons than during the mornings. Knowing that Francesc Macià is located uptown and close to offices and workplaces, it would be reasonable that this station had larger activity during the afternoons and indeed it has. If we check the real activity for this station (depicted in red at Figure 24) it can be seen that the activity occurs at 16:00 and afterwards, while the activity during the morning or the nights is low. This station shows the opposite behaviour than Station 232, where the activity was higher in the mornings as it is located down-town, close to Paral·lel, an important Metro station.

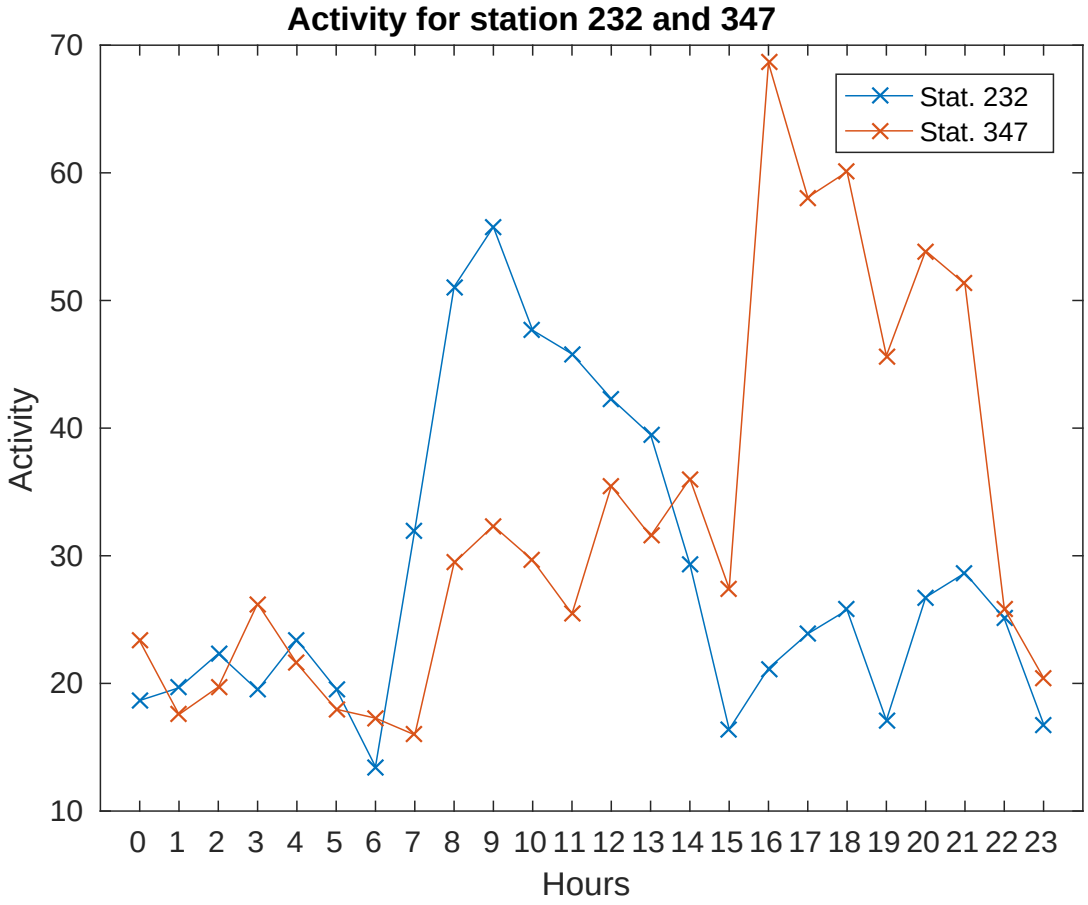

Figure 24: Activity at Stations 232 and 347

To wrap up, plotting over the principal components is interesting in order to gain visibility on the stations general behaviour. To do that, it is important to understand what behaviour each principal component explains. Moreover, once we have understood the meaning of each component, plotting the data over these components helps us to see the relationship between them in a more particular way, selecting meaningful stations and comparing the Principal Component Analysis results with the real behaviour of the stations.

3.2.2 Spatial Analysis

In this case, the Spatial Analysis consists on considering as variables the Bicing stations. Consequently, the dimension of the Covariance matrix built is 419×419 and the principal components are vectors whose dimension is equal to 1×419.

As the approach changed, the dimensions of the vectors and matrices have also changed according to the variables we have considered now, which are the stations instead of the hours. The results for Spatial Analysis are to be analysed in the Appendix, as they do not provide a different and enlightening outcome. The achieved results are similar to the previous section.

Nevertheless, it seems interesting to show here the First Principal Component, so we can see the main feature of this analysis, which is the activity over the stations (and is similar to the one obtained in the Temporal Analysis)

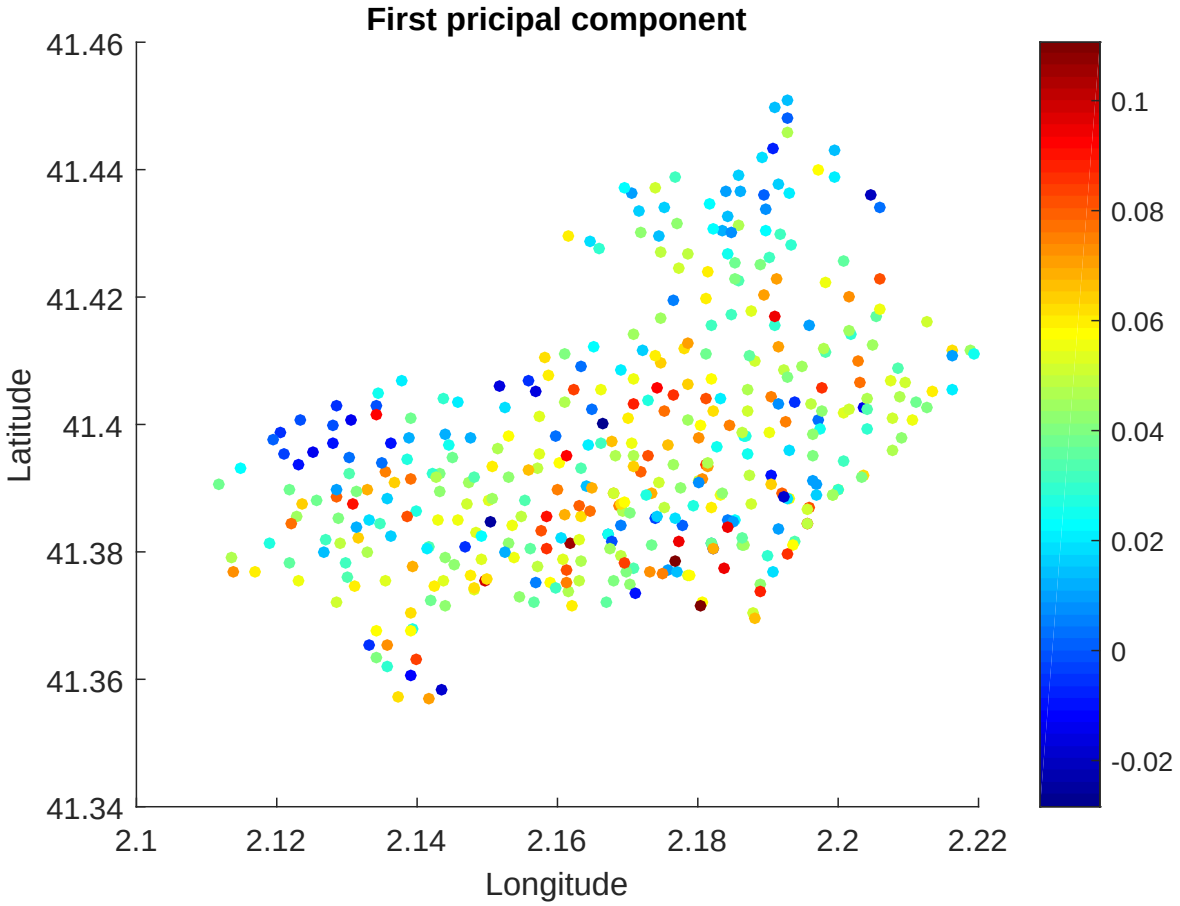

Figure 25: First Principal Component for the Spatial Analysis

4 CONCLUSIONS

First of all, it is worth mentioning the large amount of possibilities that this new era of technology, and consequently Big Data, offers in terms of discovering knowledge out of data. Finding data value can be used afterwards to improve decisions and competitiveness either for companies or public administrations and it will create a significant growth of the world economy. Thus, Big Data is the future.

Secondly, it is of great need to understand the challenges of working with such data. A Big Data dataset can have hundreds of columns and millions of rows containing information registered during years or even decades. These might imply storage problems and high computational cost. In other words, to carry out a project involving Big Data it is necessary to dispose of the required means to produce results in reasonable time. Furthermore, it needs to be pointed out that the most important part of a Big Data project is the treatment of the data before analysis. From data acquisition until data analysis, there are many steps that cannot be disregarded as they can compromise the results and all the work would be useless.

Before starting with the conclusions on the results, it is interesting to make clear that the analysis has been carried out considering the standard deviation of the number of bikes as proxy for activity. This analysis might have also been carried out considering average value of bikes, or making different data aggregations than the ones presented on this thesis. However, for the purpose this project has, which is discovering mobility patterns, this approach is more reasonable than any other as it has been possible to group the data by hours, days and moths and in consequence, find out the desired mobility patterns.

Common sense told us that the activity would be higher when the users enter or leave their workplaces, universities or schools, or that the activity would be higher on weekdays than weekends. However, common sense does not always is right. With this project, we have been able to find out when the peaks of the activity take place based on data for a whole year. Early in the mornings, weekdays and summer are the periods where the activity has proven to be higher opposite to nights, weekends and winter, where this activity is lower. Moreover, the results show the variability of the data as well, making clear that not all the days have the same behaviour but showing that they follow a common pattern.

On the other hand, Principal Component Analysis has proven that the dimensions of the data can be reduced and with only three dimensions it is possible to explain the 75% of the data variability. With only three principal components, we have been able to explain the different behaviours during a day, and show the general trends for the 419 stations. Principal Component Analysis offers more possibilities such as data reconstruction using less dimensions. However, carrying out a reconstruction of the original activity does not provide new insights in terms of mobility patterns thus it is not necessary for this project. Nevertheless, it could be a possibility in terms of future research.

Furthermore, by using Principal Component Analysis, it has been possible to identify some stations which had very high (or low) values when projected over the principal components. Comparing its position with respect to the principal components with the real activity, we have seen that Principal Component Analysis gives good results and explains the real behaviour of the stations using always less dimensions than the initial dataset.

With the present work, we have been able to determine the peak and off-peak periods for different hours, days and months. These results might be applied in terms of improving the system efficiency, providing more bikes in areas where the activity is higher in detriment of those areas where the bikes usage is inferior. Moreover, this work sets the basis for a future research using Bicing data, as the data is at our disposition, and has been pre-processed and treated. Consequently, further studies might be rather focused on data analysis than on data acquisition and treatment.

Bicing data offers a lot of possibilities in terms of data analysis. This project was focused on finding out mobility patterns using Big Data, which has been successfully carried out. However, using Principal Component Analysis, it would be possible to reconstruct the data and prove that the reconstructed results are similar than the original ones. Furthermore, using machine learning algorithms, such as neuronal networking or genetic algorithms, it would be possible to create a prediction model, able to foresee the occupation of the stations, also considering holidays, weather data and other factors that can have a repercussion on the system service.

References

[1] U. Fayad, G. Piatesky-Shapiro, and P. Smyth. From Dara Mining to Knowledge Discovery in Databases. *Artificial Intelligence Magazine*, 17(3), FALL 1996.

[2] Dough Laney. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Technical report, META Group, February 2001.

[3] P. Thakuriah, N. Tilahun, and M. Zellner. Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery. *Proc. of NSF Workshop on Big Data and Urban Informatics,*, 2015.

[4] C.F. Barnes, H. Fritz, and Yoo J. Hurricane Disaster Assessments With Image-Driven Data Mining in High-Resolution Satellite Imagery. *IEEE Transactions On Geoscience and Remote Sensing*, 45(6), June 2007.

[5] McKinsey Global Institute. Big data: The next frontier for inovation, competition, and productivity. Technical report, McKinsey & Company, June 2011.

[6] G. Martins, B. Bellalta, and S. Oechsner. Predicting Occupancy Trends in Barcelona's Bicyvle Service Stations Using Open Data. *SAI Intelligent Systems Conference*, November 2015.

[7] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, Aberdeen, U.K., 2nd edition, April 2002.

[8] Models for Decision Making and Optimization in Engineering. Notes on Principal Component Analysis. *Civil Engineering Master Course Notes - UPC*, 2015.

[9] Observatori Fabra. Resum Mensual de Temperatures i Precipitacions del 2015. http://www.fabra.cat/meteo/resums/resums2015.html.

Appendices

A Spatial Analysis

First of all, Figure 26 depicts the cumulative sum of the explained variance. In the previous section, we had only 24 eigenvalues but now we have 419 (which corresponds with the number of stations considered for the analysis). In this case, three principal components explain approximately the 75% of the data. In other words, using only three components (instead of 419) we can obtain results that explain the 75% of the variability of the data, achieving a huge dimensionality reduction. It is worth to mention again that in Big Data problems is really interesting to achieve dimensionality reduction as the initial amount of data is really large and uses several computational resources.

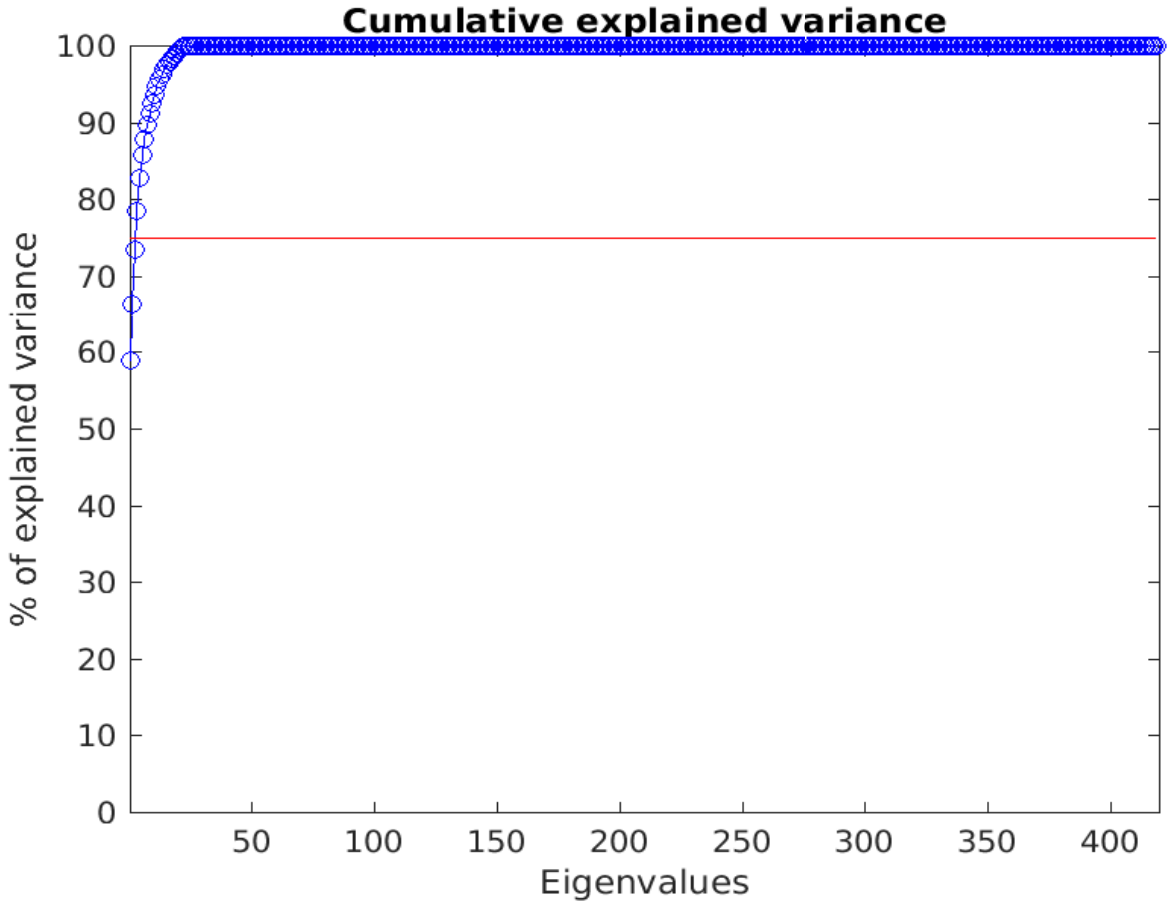

Figure 26: Cumulative sum of the explained variance by the principal components for the spatial analysis

Since the dimensions changed, if we want to take a look at the three first principal components, as we did in the previous section, we need to analyse the projections. Figure 27 depicts the principal components projected.

In this case, and similarly to what we had, first principal component projection (depicted in Blue in Figure 27) stands for the high activity intervals during a day. The different peaks of a day are ones

with highest values while the negative peaks (or the low activity hours) are also shown with small values. To easy remember what represents this projection, we can label it again as "High-Demand".

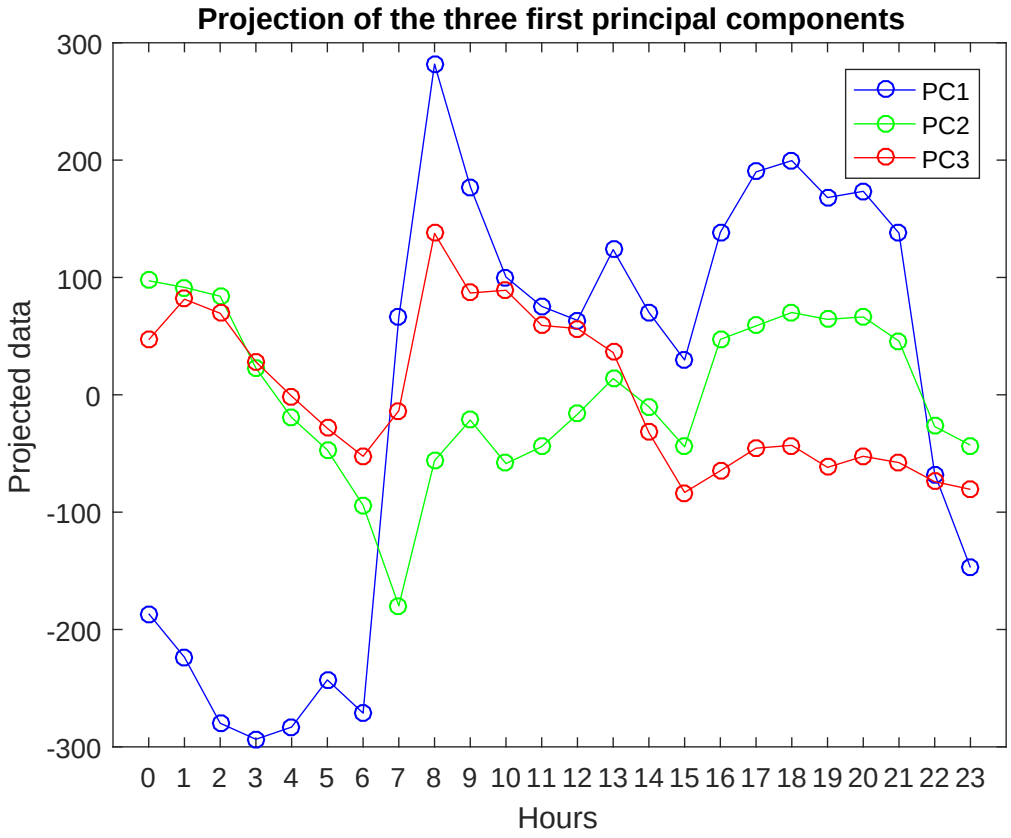

Figure 27: Projection of the three first principal components for the spatial analysis

Second principal component projected, shown in Green at Figure 27, is this time less clear than before. While the second principal component shown at Figure 13 was labelled as "Night vs. Day", the projection depicted at Figure 27 seems to explain the relationship between activity during the afternoon and the activity during the night. These two day periods are correlated as both ar positive although they have different activity patterns. Moreover, they both seem to be decorrelated to the morning peak. For purposes of making easier to understand what this projection explains, it has been labelled as "Night/Afternoon vs. Morning".

Right now, with only two components we have been able to explain the "High-Demand" periods (and consequently the low demand ones) and the relation with the activity during the day. In terms of visualization, knowing that each graph explains the aforementioned patterns, it is possible to have an idea of the activity during a day (based on information of an entire year).

On the other hand, the projection of the third principal component (depicted in Red at Figure 27) is also different to the one obtained in the previous section. Now, it seems that the projection of this principal component stands for the relationship between night and morning activity in detriment of the afternoon activity, that appears now to be decorrelated. There are two different peaks, explaining peaks of activity during the night and during the morning. On the other hand, the afternoon area seems more plain, as the activity is much more moderated than on the other periods. Once again, this projection can be relabelled but this time it will be called "Night/Morning vs. Afternoon".

Again with three components it has been possible to obtain an idea of the behaviour of the system. Furthermore, to complete this interpretation of the principal components, it is interesting to plot the data using Bicing data over First and Second Principal Components these components as a base. Figures 28, 29 and 30 represent the aforementioned approach.

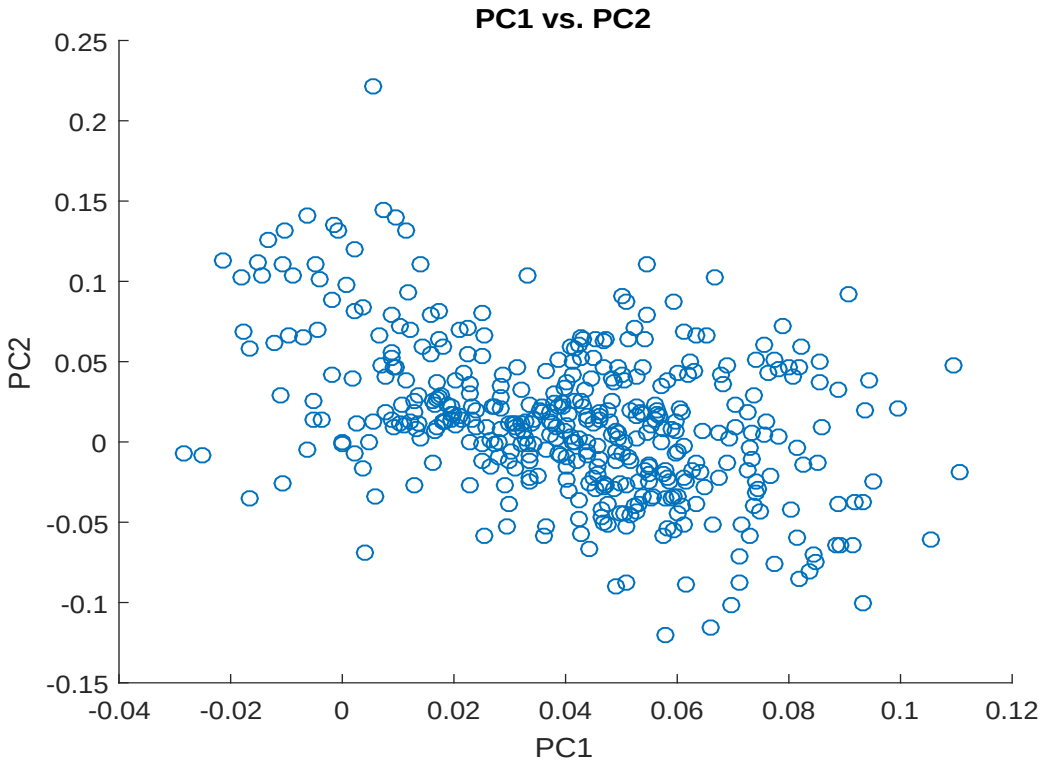

Figure 28: Bicing data over First and Second Principal Components

As before, PC1 explains the activity hours, so the dots (or stations) located to the right (on Figure 28,)are the ones with higher activity. On the other hand, PC2 explains the relationship between the behaviour during night and afternoon which is decorrelated with the behaviour during the morning. In here, the vast majority of dots are located on the centre of the vertical axes, which means that they do not behave very different during the afternoon or the morning. Furthermore, the stations located on the extremes are those that have more activity during night/afternoon than during the morning (upper part) and the other way around, depicted in the lower part of the scatter plot. As it can be seen, the tendency is to stay in de centre and closer to the negative area, explaining that most of the stations have higher activities during the morning when everybody uses Bicing to go to work, university or school.

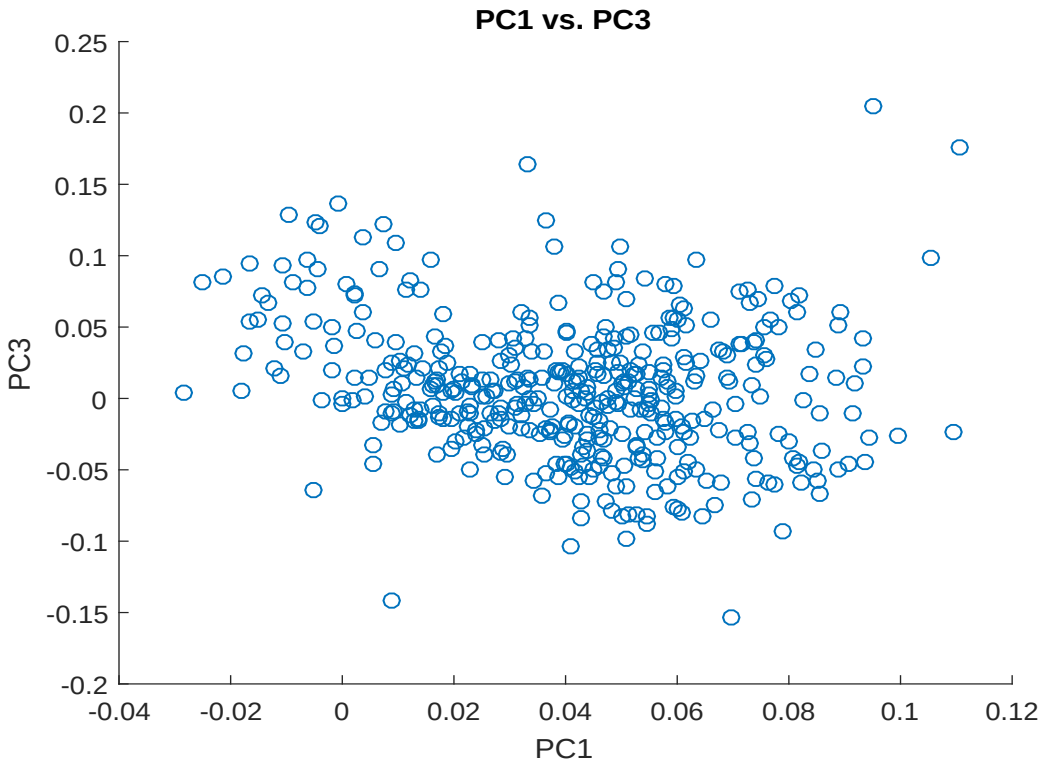

Figure 29: Bicing data over First and Third Principal Components

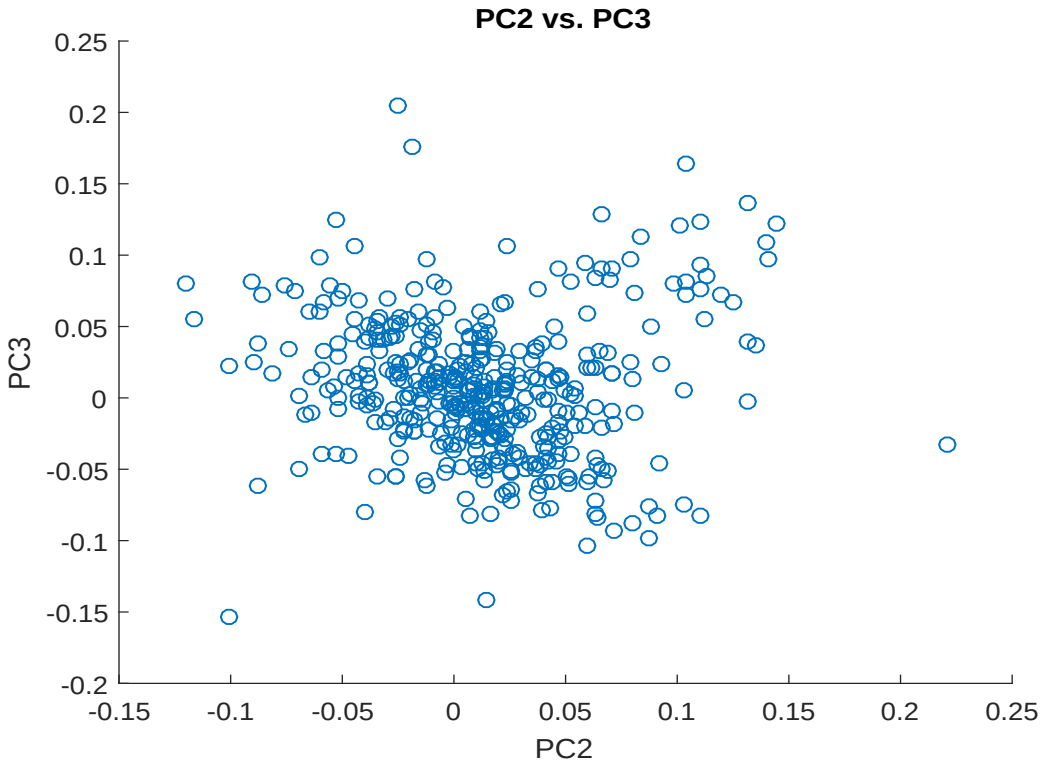

Figure 30: Bicing data over Second and Third Principal Components

Figure 29 has PC3 as vertical axis and the distribution of the dots is quite similar. The tendency of the stations is once again to stay on the central part. The extreme dots represent now those stations that have more activity during the night and morning than the afternoon (upper part) and those with the opposite behaviour (lower part).

On the other hand, Figure 30 represents the data using first and third principal components as a base. As it has been mentioned, only few stations present differentiated behaviour during the day periods. Consequently, almost all the dots are located around the point (0,0). If we take a look at the dots on the upper-right corner, these are stations with large values of PC2 and PC3, which means that they have a large different between the activities during a day. The same happens on he opposite corner (lower-left).